정보 서비스를 위한
분류, 시소러스, 온톨로지의 이해

서태설, 김비연

한국과학기술정보연구원

<발 간 사>

정보 서비스에 있어서 분류는 핵심 도구 중의 하나입니다. 그런데, 최근 정보 서비스 환경의 변화에 따라 분류뿐만 아니라 시소러스나 온톨로지도 중요하게 인식되고 있습니다. 분류, 시소러스, 온톨로지는 오늘날 지식정보를 다루는데 활용되는 핵심 도구이지만, 각각의 의미와 역할에 대해서 사람마다 조금씩 다르게 이해하거나 혼동하기 쉬운 것이 사실입니다. 실제로, 분류에는 계층적 분류와 트리구조의 분류로 세분화되고, 시소러스와 온톨로지도 만드는 주체에 따라서 동일한 대상에 대해서 다르게 구성될 수가 있기 때문에 활요할 때 이러한 배경 지식을 충분히 이해한 후 적용해야 합니다.

따라서, 본서에서는 정보 관리와 서비스에서 바라보는 분류, 시소러스, 온톨로지에 대해서 이론적인 배경을 살펴보고, 실제 사례를 소개함으로써, 연구자나 실무자가 이러한 문제에 직면했을 때, 조금이나마 도움이 되도록 하였습니다.

먼저, 분류의 경우 관련 용어의 개념을 다룬 후, 분류 체계(classification scheme)의 유형과 요건, 선정 기준 등을 제시하였으며, 시소러스(thesaurus)와 온톨로지(ontology)도 정의, 목적, 핵심 요소, 개발 방법 등을 소개하였고, 마지막으로, 각각의 주요 구축 사례를 소개하였습니다.

모쪼록 본서가 국내 정보 서비스 발전에 조금이나마 기여할 수 있기를 간절히 바라마지않습니다.

2015년 10월

한국과학기술정보연구원장
한 선 화

목 차

1. 서론

1.1 분류 개요

1.2 개념

1.3 범주화

1.4 지식 및 학문분류

1. 서론

1.1 분류 개요

최근 언어학, 전산학, 문헌정보학 등의 연구분야 뿐만 아니라 많은 조직에서 어휘집, 시소러스, 의미망, 온톨로지, 텍사노미 등으로 표현되는 지식 정보의 체계적인 구조 확립에 대한 논의와 연구들이 진행되어 왔다. 이들은 모두 개념이나 어휘들의 의미적 계층구조 형성과 의미관계 설정에 관한 공통적인 관심사를 가지고 있으나 분야에 따라 표현 및 설명의 차이가 나타나고 있어 여전히 논쟁의 여지가 많고 합의된 정의를 제시하기가 용이하지 않다.

분류의 영어표기는 'classification'과 'taxonomy'로 전자는 '나누다', '정리하다'라는 의미로 자료분류에서 범용되고, 후자는 '나누다'는 의미로 주로 동식물분류를 위한 계층분류시스템에서 비롯되었다. 문헌정보학에서 전통적으로 사용해 온 분류라는 용어는 지식의 구조단위인 개념과 이러한 개념간의 관계와 범주화를 통해 특정 영역의 지식을 일정한 체계로 구조화하는 의미로 사용되어 왔다. 오늘날 텍사노미는 객체를 기술하거나 검색을 용이하게 하는 방법으로 주제 기반에 의한 용어들을 계층적으로 범주화하는 것과 관련된 접근방법으로 널리 사용되고 있다(Garshol, 2004).

분류나 텍사노미는 모두 기본 개념이 '분류학'에서 비롯되었으며 용어의 표현과 사용방법이 분야에 따라 차이가 있지만 내용적으로 공통된 특성이 많기 때문에 분류와 텍사노미를 동일한 맥락에서 다루더라도 의미상 큰 문제가 없다. 단지 텍사노미의 경우 분류체계가 계층적 구조를 따른다는 점에서 넓은 의미의 분류 안에 포함된다고 할 수 있다.

분류는 주로 지식을 체계화하고 조직화하는데 주된 관심이 있으며, 기호로 표현된 지식을 일정한 구조체계 아래 조직하기 위한 이론과 기법에 관한 영역이라고 할 수 있다. 그런데 분류는 유용성이 전제되어야 하기 때문에 실천적이며 도구적이어야 한다. 따라서 분류에서는 보다 효과적인 방식으로 개념을 구조화하는데 본질적인 관심을 두고 있으며, 이를 위해서는 지식 자체의 성질과 구조의 이해가 선행되어야 한다(김태수 2000, 9).

분류와 더불어 시소러스와 온톨로지는 모두 지식을 어떻게 체계화시키고 지식을 어떻게 저장하며, 저장된 지식을 어떻게 활용할 것인가에 대한 연구라고 할 수 있으며 모두 주제나 개념을 범주화하는데 본질적인 관심을 두고 있다. 따라서 기본적

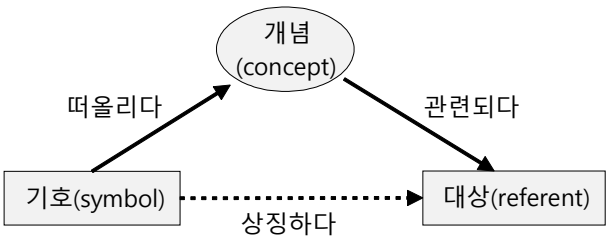
으로 지식을 표현하고 조직하기 위한 분류의 기본도구인 개념과 범주에 대한 이해가 요구된다.

1.2 개념

일반적으로 개념은 동일 속성을 지닌 대상물로부터 추상하여 일반화된 관념으로 정의할 수 있다. ISO(1987)에서는 '사고의 단위 요소' 혹은 '외부 세계와 내부 세계에 존재하는 개개의 대상을 다소 임의적으로 추상하여 분류한 정신적 구성물'로 정의하고 있다.

개념의 특성은 일반적으로 특정 언어나 단어에 종속되어 있지 않다는 것이다. 개념은 사물에 대한 지적 또는 논리적 표현으로서, 이를 상징하는 기호(symbol)가 반드시 필요한 것은 아니지만 기호 없이는 커뮤니케이션이 불가능하므로 대부분의 개념에는 기호가 부여된다. 이러한 과정에서 같은 개념에 여러 기호가 부여되기도 하고 다른 개념에 같은 기호가 부여되기도 한다.

기호를 이용한 커뮤니케이션 과정은 '의미의 삼각형(meaning triangle)'으로 설명될 수 있는데 사물과 개념, 기호 간의 상호작용을 설명하는 이론이다. <그림 1>은 기호가 나타내려고 하는 본질(개념)과 그 기호가 상징하는 지시 대상(사물) 간의 삼각관계를 나타낸다. 즉, 인간은 특정 기호(단어)를 보고 그 기호가 상징하는 사물과 연결할 때, 그 기호와 어떤 사물을 직접적으로 연결하는 것이 아니라 그 기호와 일치하는 개념을 먼저 떠올리고 그 개념과 상응하는 사물과 관련시키는 인지과정을 통해 커뮤니케이션을 이루게 된다.



<그림 1-1> 의미의 삼각관계

개념의 또 다른 특성은 '내포'와 '외연'이라는 두 가지 속성을 가지고 있다는 것이다. 내포(intension)는 '의미'를 말한다. 이는 '내부에 포함하여 가짐'이라는 의미로 논리학에서는 어떤 개념의 내용이 되는 여러 속성을 일컫는다. 외연(extension)'이

란 어떤 개념이 적용되는 명제나 사물의 범위를 일컫는 말이다. 예를 들어 '동물'이라는 개념의 외연에는 새, 물고기, 사자, 인간 등이 있다(노상규, 박진수 2007, 22-23). 따라서 포괄적인 개념은 다양한 유형의 대상을 그 의미 범위에 포함시키기 때문에 외연이 넓은 반면, 다수의 특성을 지닌 개념이 지시하는 대상은 실세계에서 소수로 제한된다.

개념은 원칙적으로 하나 이상의 외연적 대상 수와 내포적 속성을 지닌다. 그런데 특성 자체도 개념이기 때문에 개념의 특성에 따라 계층구조를 표현할 수 있다. 이러한 개념 사다리에서는 상위 개념이 지닌 특성이 하위 개념으로 계승되고 최하위 수준의 개념은 전체 개념의 속성을 소유하게 된다. 이런 까닭으로 개념간의 계층을 인정할 수 있으며 개념은 특성에 의해 구조화될 수 있다(Dahlberg 1978, 145)

개념을 통해 특정 단어나 사물을 단순히 보고 듣는 것 이상의 정보를 추론할 수 있는데 개념의 기능을 요약하면 다음과 같다(김태수 2000, 15).

첫째, 인지적 경제성과 안정성을 얻을 수 있다. 현실 세계를 몇 개의 유목으로 범주화하면 지각하고 기억하고 추리하여 의사소통과정에 필요한 정보량을 크게 줄일 수 있다.

둘째, 개념이 지닌 정보 이상을 알 수 있다. 지각 정보를 비지각 정보와 연결시키는 도구로서의 기능과 이미 알고 있는 사실에서 논리적인 결론을 추론할 수 있는 기능을 일컫는다.

셋째, 개념의 조합을 통해 복합적인 개념이나 사고를 형성할 수 있다.

1.3 범주화

분류 뿐만 아니라 여러 학문영역에서도 범주화에 대한 논의가 중요한 관심사가 되어 왔는데 개념과 범주, 이들의 구조와 조직이 지식의 구조화에 핵심요소이기 때문이다. 특히 범주화는 인간활동의 근본으로서 다양성 속에서 유사성을 파악하기 위한 인간의 능력이다(Taylor 1997, vi-viii). 다시 말해 범주화란 인간이 현상 세계를 의미 있는 단위로 나누어 파악하는 장치이다(임지룡 1997, 92)

개념을 범주화하게 되면 이 범주에 속하는 개념은 상호 관련을 가지게 되고 공통의 특성을 지니게 된다. 범주화를 통해서 얻을 수 있는 유용성은 1) 환경과 상호작용이 가능하며 환경의 복잡성을 줄일 수 있으며, 2) 사물의 유형을 쉽게 확인하고 인식할 수 있고, 3) 계속적인 학습의 필요성을 감소시키며, 4) 적절한 행위를 결정할 수 있고, 5) 물건이나 사상을 분류할 수 있다는 점을 들 수 있다.

개념의 범주화 방법은 관점에 따라 다른 기준이 적용될 수 있다. 물리적 개체나 추론기법, 분석기법, 성질에 따라 개념이 범주화될 수 있으며 사용 목적에 따라 구체적인 대상이나 추상적인 대상, 제품, 제조 과정, 기타 유용한 성질을 기준으로 범주화될 수도 있다. 또한 특수 주제 분야의 개념은 그 주제 영역에서 사용되는 개념의 성질에 의해서도 구조화될 수 있다. 분류론의 기초이론으로서 범주화 모형의 유형과 시대적 변화의 특징을 살펴보면 다음과 같다(김태수 2000, 27-52):

가장 대표적인 범주화 이론은 고전범주화모형으로 고전적이라는 것은 고대그리스의 아리스토텔레스의 기본 범주까지 거슬러 올라간다는 것과 20세기 줄곧 심리학과 철학, 언어학 등을 지배해 왔다는 것을 의미한다.

일반적으로 고전범주화모형이 주는 의미는 범주는 필요충분 자질의 집합으로 범주 내의 모든 성원은 공통의 속성을 지니며 그래서 모든 성원은 대등하다는 점이다. 이 범주에서는 경계가 분명하고 이 경계는 특성의 정의로 결정된다.

전통적인 계층 분류에서는 고전범주화이론을 도입하여 개념간의 관계를 불변의 관계로 전제로 한다. 이 분류법에서는 개개의 지식 영역을 분명한 경계로 구분하여 독립된 주류(main class)로 구분하며 모든 주제를 이 주류에 포함하였다. 이 분류법에서는 특성을 공유한 개념들의 범주를 결정하기 위한 정의와 전체 지식 구조 내에서 개념간의 관계를 체계적으로 표현하기 위한 배열의 문제를 분류의 주된 기능으로 하고 있다. 이 전통적인 분류법은 지식 구조에 분명한 경계를 인정함으로써 유연성이 적고 폐쇄적이어서 각 학문영역의 지식 발전을 수용하는데 한계로 지적되고 있다.

반면 특정 개인의 실세계의 경험 지식에 대한 구조화 또는 범주에 포함되는 성원의 위계 구조의 존재와 범주화의 생태학적 기반이 인정되면서 자연범주화에 관심이 집중되었다. 자연범주화의 관점은 범주는 본질적으로 불안정하고 항상 상황에 따라 지속적으로 재구성되면 목표 지향적이므로 범주를 분석적으로 기술할 때 연속적으로 변화하는 현상을 구조에 반영할 수 있도록 해야 한다는 것이다. 달리 말하면 사물의 속성은 사물 자체의 내재적 속성과 관계가 있는 것이 아니라 특정 문화권에서의 사물의 역할과 관계가 있다는 것이다.

자연범주화에서는 특정 범주에서 가장 전형적이고 대표적인 성원의 존재를 인정하는데 가장 전형적인 성원을 원형(prototype)이라고 하며 이 원형은 기본적으로 문화와 환경에 따라 달라진다. 원형이론이 가지는 일반적인 성질은 범주의 성원을 결정하는데 필요하고도 충분한 속성에 대한 제한이 없다는 것, 범주의 경계가 불분명하여 어떤 성원은 다른 범주에 포함될 수 있다는 것 등을 들 수 있다. 자연범주

의 다른 특성은 범주의 성원을 가장 전형적이고 대표적인 것에서부터 비전형적인 성원에 이르기까지 연속적으로 배열할 수 있는데 이를 위계구조라고 한다. 위계구조에서 범주를 정의하고 성원의 순위를 결정하는 것은 누가 정의하느냐에 달려 있다.

일반적으로 지식의 전달과정에서는 실용성이 중요시되며 개념을 쉽게 이해하여 이를 생각과 행동에 이용하고자 한다. 따라서 자연범주에서 제시하는 모형이 이러한 목적에 더 적합하다는 것이다. 따라서 모든 지식을 조직할 수 있는 일반적인 표준분류법을 거부하고, 개개의 분류표는 자체의 조직구조와 고유한 특성을 지녀야 한다는 것이다. 인간의 사고 패턴이 하나가 아니듯이 모든 상황에 적용할 수 있는 포괄적인 분류체계는 있을 수 없다는 것이다.

자료 분류에서는 고전적인 범주화의 기법을 일찍이 도입한 바 있다. 의미관계라는 관점에서 보면 카드 목록 시대에 이미 상호 참조를 사용하여 개념간의 의미 구조를 표현하였다. 그런데 1940년대에 와서 지식의 구조화와 관련하여 개념간의 연결기법이 제기되었다. 즉, Bush는 상이한 개체간에 연결할 수 있는 메믹스(memex) 시스템을 제안하였는데 이 시스템에서 개체간의 연결은 비합리적이고 비논리적이며 일관성이 없다는 점이다. 하나의 생각과 다른 생각, 하나의 용어와 다른 용어간의 관계에 대해 마음속에서 이루어지는 연상은 특이한 양태를 보인다는 것이다. 따라서 하나의 지식 조각에서 다른 지식 조각으로 연결할 수 있는 이른바, 지식의 통로를 제안하였다. 이 통로가 바로 오늘날의 하이퍼텍스트와 웹의 선구자인 것이다. Lesk는 이 메믹스 개념은 지식의 구조화에 인간의 인지 과정을 고려하여 개념과 개념을 연결한다는 점에서 그 의미를 부여하였다.

한편 1950년대에 이후부터 인지 모형을 반영한 분류론이 제기된 바 있다. 즉 인간의 뇌와 같이 감각을 통해 다수의 현상을 관찰하고 이를 다른 현상과 비교하여 다수의 복잡한 미지의 사물을 한정하고 관련지어 사물간의 경향이나 상호의존성, 관계를 설정하는 것이다. 특히 Soergel은 색인과 카드 목록에서 개념간의 연결기법을 제안하여 지식의 구조화와 정보 검색에서 비계층 관계의 적용가능성을 제기한 바 있다. 즉, 색인자나 탐색자가 개념 A를 사용할 때 개념 B의 존재를 연상하게 되면, 그리고 이 두 개념이 계층 관계가 아니라면 개념 A는 개념 B와 관련을 가진다는 것이다. 이러한 의미 관계를 가진 개념 간을 연결하는 것이 검색의 효율을 개선할 것이라고 지적한 바 있다.

한편 Jacob은 분류에서 전통적인 범주화 이론뿐만 아니라 자연범주화 모형을 함께 고려할 것을 제안하였는데 그 이유를 다음과 같이 제시한 바 있다. 즉 범주화

에서는 개념이 지닌 속성이나 포섭관계만이 범주의 성원을 결정하는 유일한 기준이 될 수 없으며 개인의 인지과정과 사물 간에 존재하는 기능적 관계를 이해하는 것이 중요하다고 지적하였다.

1.4 지식 및 학문 분류

철학에서 인공두뇌학에 이르기까지 오랫동안 학자들은 많은 분야에서 지식의 개념과 정보 시스템 안에서 지식을 표현하는 문제를 논의해 왔다. 일반적인 의미에서 지식이란 유전이나 경험을 통해 학습된 모든 정보이며, 보다 구체적이고 학술적인 표현으로 바꾸어 서술하면 지식이란 “세상에 대한 이해를 위하여 개인, 사회의 여러 그룹이나 조직, 문화권, 또는 인류 전체에 존재하는 구체적이고 추상적인 대상과 사건 그리고 사실 내용에 관한 모형들의 현재”라 할 수 있다. 통시적 관점에서 볼 때 담론적 또는 학술적 지식이 사회의 주류로 자리 잡게 된 것은 근대의 후기 이후에 나타나는 현상이며, 그 이전에는 다른 형태의 지식 모형이 행동의 배경으로 작용하였다(고영만 2005, 11).

그런데 이 지식은 절대적인 구조를 가진 개체가 아니라 관련 영역의 전문가들이 인식하고 있는 현재의 지식 상태를 반영한 것으로 다양하게 연결된 개념으로 구성되어 있다. 특정 시점의 지식 구조를 확정짓는 것은 어려우며 지식 체계는 항상 변화하고 유동적이다.

지식을 조직해야 하는 이유는 단순한 경험 수준이 아니라 정보를 경제적으로 획득하고 저장하여 이를 통해 현실 세계에 존재하는 사물의 행동 양식을 이해하기 위한 것이다. 이 지식은 일상의 행동을 지시하는 인간 행동의 기초로서 사물을 구분하고 분류하는 일은 일상생활에서는 물론 학문에서도 기본적이고 필수적이다. 지식은 자연 언어에 의한 구문 구조 및 의미 구조와 관련을 지니며 데이터베이스에서도 레코드나 파일 구조를 통해 지식을 표현하며 인공지능 등 다양한 분야에서 지식표현과 처리기법을 다루고 있다(김태수 2000, 53)

지식과 동의어로 사용되기도 하는 학문영역의 분류는 현대 학문의 뿌리를 형성시킨 아리스토텔레스의 분류법에서 비롯되었으며, 이후 수많은 학자들에 의해 인간의 사유를 조직화하려는 시도가 이어져 왔다. 아리스토텔레스는 학문을 목적에 따라 이론학(인간의 지식행위와 상관없이 자연적으로 존재하는 것으로 신학, 수학 및 자연과학), 실천학(정치학, 경제학, 윤리학으로 합리적 행위의 실천을 위한 지식), 제작학(예술과 기술, 응용과학으로 인간에게 필요한 물품의 제작을 위한 지식)으로 구분하였다.

16세기 이후 베이컨은 학문의 존재 양식 관점에서 기억(지리와 역사로 자연의 역사, 세속 사회의 역사, 교회의 역사 등), 상상(예술과 문학으로 서사시, 극시, 풍자시 등), 이성(철학으로 제1철학 또는 여러 학문의 원천에 대한 철학, 신, 자연, 인간 등에 관한 것으로 기억과 상상에서 다루는 주제 이외의 모든 주제)으로 학문의 세계를 나누었으며 이는 이후 자료 분류의 기초가 되는 해리스의 분류에 많은 영향을 미쳤다. 학문의 분류는 베이컨 이후에도 데카르트, 스펜서, 헤겔, 꽁트, 맑스 등을 거치면서 문화적 특성과 학파에 따라 다양한 관점에서 분류되었으며, 오늘날에는 학문의 목적, 학문의 대상과 존재 양식 외에도 학문의 연구 방법, 실용성 등에 따라 분류가 이루어지고 있다.

학문 분류는 근세까지 철학적 의미와 주요 학문 명칭을 부여하는 것에 중점을 두고 발전되었으나 현대에 와서는 각 나라에서 다양한 학문 분야의 연구 개발 지원, 측정 및 분석을 위해 실용적으로 활용하기 위한 연구 분야를 바탕으로 고안한 것이 주를 이루고 있다. (고영만 2005, 25)

2. 분류 관련 용어의 이해

2.1 분류의 일반적 정의

2.2 자료 분류의 정의

2.3 어휘, 텍사노미, 시소러스 및 온톨로지

2. 분류 관련 용어의 이해

2.1 분류의 일반적 정의

일반적으로 분류는 서로 유사한 것 끼리 그룹화 시키는 것을 의미한다. 다시 말하면 동일한 사물의 대상을 한 곳에 같이 모으고, 동일하지 않은 사물을 서로 분리시키는 것이다. 분류의 사전적 의미를 찾아보면 다음과 같다.

"어떤 대상을 공통된 특성 또는 인지되거나 추정되는 관련성에 따라 다수의 서로 구별되는 유별에 체계적으로 분배, 할당, 배열하는 행위 또는 그 결과" (Oxford English Dictionary, 2015)

"개념의 외연을 정확하게 구분함으로서 완전한 체계를 조직하는 것 즉, 사물의 대상을 최고의 유개념으로부터 최저의 종개념을 철저히 분석해 내는 것"(철학대사전, 1963, 563)

"개념이나 대상을 인지하고, 차별화하고 이해하는 과정" (Wikipedia, 2015)

이외에도 윤희윤은 분류의 정의에 대해 포괄적으로 정리를 하였는데 그는 "분류의 명사적 정의는 사전에 결정된 일련의 원칙에 따라 순서화한 범주 시스템으로서, 사례나 실체를 하나의 세트로 조직하는데 사용되는 수단이다. 동사적 개념은 하나의 분류 시스템에서 사례나 실체를 범주화하는 과정이다. 한편, 분류의 일상적 개념은 사물을 종류별로 구분하는 것이며, 학문적 접근에서는 논리학에서의 정의처럼 어떤 사물이나 대상을 일정한 기준(성격, 특징 등)에 따라 상위의 유개념에서 하위의 종개념까지 체계화는 과정"이라고 하였다(2013, 5).

Kwasnik는 "분류란 경험을 의미있게 군집화(clustering)하는 것"이라고 하고 분류과정은 탐색적 도구로서 질의의 사전 단계로 유용하며 발견, 분석, 이론화를 형성하는 중요한 방식으로 적용될 수 있다. 일단 개념이 구체화되고 개념들 간의 관계가 이해가 되면, 분류는 알고 있는 것에 대해 풍부한 표현의 수단으로 사용될 수 있고, 커뮤니케이션과 탐색, 비교, 이론화의 새로운 주기(cycle)를 만들어 내는데 있어 유용하다고 하였다(1999, 24).

통상적으로 분류와 관련하여 구분(division)이 같은 의미로 혼용되지만 엄밀하게는 더욱 특수한 것에서 다음의 추상적으로 나아가는 과정 즉, 종합적·귀납적인 방법을 분류라고 하고 그 반대로 일반적인 것에서 특수한 것으로 나아가는 과정 즉, 분석적·연역적인 방법을 구분이라고 한다. 따라서 분류란 상승에 의한 종합적인 방법으로 최저의 종개념에서부터 최고의 유개념에 도달하는 것을 의미하고, 구분은 하

강에 의한 분석적인 방법으로 특정한 유개념을 분석하여 최저의 종개념에 도달하는 것이라고 할 수 있다(김포옥, 백항기, 2011, 27).

2.2 자료 분류의 정의

문헌정보학 용어사전에 의하면 자료 분류란 “사물이나 현상, 개념 등을 유사한 것은 모으고, 상이한 것은 구분하여 체계화하고, 그 결과 분류된 사상의 명칭이 체계적으로 배열된 표”로 정의된다(2000, 166). 자료 분류에서는 이러한 표를 통해 개념을 구체적으로 조직하고 있으며 이 체계표를 조직하는 방법에는 두 가지가 있다. 하나는 귀납적인 방법으로서 공통의 특성을 지닌 개개의 사물이나 개념을 유사성에 따라 조직하여 점진적으로 모든 개체를 포괄한 완전한 체계를 조직하는 방법이다. 다른 하나는 연역적인 방법으로서 완전한 체계를 전제한 다음 몇 가지 기준으로 구분하여 특정한 개체에 이르는 방법이다. 이 때 개개의 분화단계를 범주라고 하고 이를 서열화한 것을 계층이라고 한다. 생물분류에서는 계, 문, 강, 목, 과, 속, 종의 7단계를 거쳐 분류하는데, 도서의 분류에서도 이 중에서 문, 강, 목과 같은 용어를 사용하며, 분류의 기준으로서 추출된 특징이나 식별점을 분류원리라고 하고, 일반적으로 공통의 원리가 모든 범주에 적용되어 계층을 형성하는 것이 바람직하다고 본다.

윤희윤은 자료 분류를 “도서관이 입수하는 모든 정보자료의 배가 위치를 결정하는 동시에 접근 이용의 편의성을 제공하기 위하여 주제나 형식의 유사성 또는 특정 원칙이나 목적에 따라 체계적으로 조직하는 행위나 과정이다. 달리 표현하여 일련의 정보 자료를 주제, 형식, 관점, 지역 등에 따라 군집화(grouping)하고, 군집된 유개념 내지 종개념 내에서 각각을 다시 개별화(individualization)하는 절차”라고 하였다(2013, 14)

자료 분류의 본질은 주제와 기타 패싯을 나타내는 기호 시스템의 논리적 조합이라는 점이다. 그것은 도서관이 분류 도구로 사용하는 모든 분류표는 주제를 논리적으로 배열한 리스트이며 각각의 리스트는 주제를 표현하는 자신의 기호를 가지고 있다는 사실에서 유추할 수 있다. 다른 하나는 지식정보 세계를 조직하는 행위이며, 이용자의 서가접근을 지원하는데 목적이 있다(윤희윤 2013, 14).

자료는 학문영역의 연구 성과를 수록한 것이기 때문에 자료 분류에서 말하는 체계화나 조직화는 바로 학문 분류를 전제로 한다. 전통적인 학문 분류에서는 일상에 적용하는 자연 분류와 달리 지식 전체를 일정한 특성과 체계에 따라 다수의 하위

학문 영역으로 구분해 왔다. 여기서 학문 영역이란 내포와 외연이 일관되게 내재된 조직적인 개념 체계로서 지식과 동의어로 사용되기도 한다(김태수 2000, 119) 자료 분류는 논리학 상의 분류 개념과 일치하고 학문 활동에서 이루어지는 학문적 분류 이론과 축을 같이 하고 있으며, 대부분의 현대적 문헌분류표가 학문 분류의 바탕 위에서 창안되고 발전되었다.

2.3 어휘, 텍사노미, 시소러스 및 온톨로지

2.3.1 통제어휘집(controlled vocabulary)

통제어휘집은 분명하게 열거된 용어의 리스트이며, 통제 어휘집에 있는 모든 어휘는 애매모호하거나 중복되지 않는 정의를 가지고 있어야 한다. 그것은 용어를 통제어휘집에 등록하는 일과 관련하여 해당 기관이 얼마나 엄격하게 통제 어휘집을 다루느냐에 달려 있으며 최소한 다음 두 가지 규칙이 적용되어야 한다:

① 만약 같은 용어가 다른 문맥에서 다른 개념으로 사용될 때에는 애매모호함을 해결하기 위해 그 명칭을 명확하게 한정(qualify)해야 한다.

② 만약 여러 개의 어휘가 동일한 의미로 사용되고 있다면, 그 중 한 용어를 우선어(preferred term)로 지정하고 나머지 용어를 동의어 또는 이칭어로 처리해야 한다.

통제 어휘집은 의미가 구체적으로 명시되지 않고 단지 사람들이 사용하기로 동의하고 그 의미가 이미 이해된 용어들의 집합일 수도 있고, 또는 각각의 용어들에 대해 매우 상세한 정의를 가지고 있을 수도 있다.

2.3.2 텍사노미(taxonomy)

텍사노미는 통제 어휘집 용어를 계층적 구조로 조직화한 컬렉션(collection)이다. 텍사노미에는 다른 유형의 부모-자식(parent-child) 관계들(전체-부분, 속-종, 유형-인스턴스 등)이 있다. 올바른 실행을 위해서는 모든 부모-자식 관계를 단일한 부모(parent)로 제한하여 같은 유형(type)이 되게 하여야 한다. 어떤 텍사노미는 복수 계층구조(poly-hierarchy)를 허용하는데 이는 하나의 용어가 여러 곳에서 복수의 부모를 가질 수도 있다는 것을 의미한다. 특별히 어떤 용어가 텍사노미 내의 한 곳에서 자식(children)을 가질 경우, 그 용어가 나타나는 모든 다른 위치에서도

같은 자식을 갖는다.

텍사노미는 계층적 링크의 의미가 무엇이건 간에 이 링크를 통해 명세화된 추가적인 의미를 갖는다. 전통적인 텍사노미의 의미는 어느 방향으로 가느냐에 따라 일반화/특수화 또는 '…는 …의 종류'를 일컫는 반면, 오늘날 텍사노미는 링크를 위해 다른 의미를 가진 여러 가지 종류의 계층 구조(예를 들어 …의 일부, …보다 넓은 주제, …의 사례)를 지칭하는 데 사용되고 있다. 텍사노미가 계층적 링크를 위해 다양하고 주의깊게 정의된 의미들을 갖는다면, 온톨로지와 아주 유사해질 것이다.

2.3.3 시소러스(thesaurus)

시소러스는 통제 어휘집 용어들이 서로 연결되어 네트워크를 이루고 있는 컬렉션(collection)이다. 이는 시소러스가 부모-자식(parent-child) 관계와 더불어 연관(associative) 관계도 사용한다는 것을 의미한다. 시소러스 내에서 연관 관계를 표현하는 방법은 다양한데 A란 용어가 B란 용어와 관련될 때 단순히 "관련어(RT)"라고 표현할 수 있다.

시소러스에는 광의어/협의어 그리고 일반화/특수화의 두 가지 용어관계는 앞서 기술한 텍사노미와 다르지 않다. 시소러스는 다른 종류의 용어 관계도 있으나 계층적 관계가 아닌 경우가 일반적이다. 물론 계층적 관계일 수도 있으나 이 연결 고리는 명백한 의미가 전혀 없고 단지 두 용어 사이에 어떤 관계가 있다는 것만 표현될 수 있다.

2.3.4 온톨로지(ontology)

사람들은 glossaries와 data dictionaries, thesauri와 taxonomies, schemas와 data models, formal ontologies와 inference 등을 의미하기 위해 온톨로지라는 단어를 사용한다. 형식 온톨로지(formal ontology)는 온톨로지 표현 언어로 표현된 통제 어휘집이다. 이 언어는 어휘 용어를 위한 문법을 갖고 있으며 특정 관심 도메인 내에서 유의미한 것을 표현할 수 있다. 이 문법은 온톨로지의 통제 어휘집에서 용어들이 어떻게 함께 사용되는 지에 대한 형식 제한을 포함하고 있다. 사람들은 관심 도메인을 위해 특정한 통제 어휘집 또는 온톨로지를 선택하여 사용한다.

이상에서 살펴본 통제 어휘집, 텍사노미, 시소러스, 온톨로지가 모두 공통적으로 갖고 있는 특징은 첫째, 어떤 커뮤니티의 관심 주제와 관련된 개념과 관계를 구축/

분류/모델링/표현하는 데 도움을 주기 위한 접근방법이다. 둘째, 한 커뮤니티 내에서 같은 용어를 같은 방식으로 사용하도록 동의하기 위해 만들어졌다. 셋째, 어떤 커뮤니티가 이 개념과 관계들을 가리키는 데 동의한 용어의 세트가 존재한다. 넷째, 용어의 의미는 어떤 방법에 의해 어느 정도로 구체화된다. 다섯째, 다른 개인들이나 커뮤니티에 따라서 많은 다른 방법으로 사용되기 때문에 혼란을 줄 수 있다.

반면 이러한 접근방법들이 구별되는 주요 차이점은 첫째, 각 용어에 대해 얼마나 많은 의미가 구체적으로 명시되어 있는가? 둘째, 의미를 명세하기 위해 어떤 표기법(notation) 또는 언어가 사용되고 있는가? 셋째, 이들은 각각 무엇을 위한 것인가? 텍소노미, 시소러스, 온톨로지는 각각 상이한, 그러나 중복되는 용도를 지니고 있다.

3. 분류 체계의 유형

3.1 구조 원리에 의한 유형

3.2 구성 방식에 의한 유형

3. 분류 체계의 유형

3.1 구조 원리에 의한 유형

3.1.1 계층 구조(hierarchies)

계층 분류의 근원은 아리스토텔레스로부터 기원한다. 그는 모든 자연을 하나의 통합된 전체로 간주하고 순서대로 체계적인 규칙에 따라 하위로 구분하는 과정을 따르고 있다. 이와 같은 아리스토텔레스의 유산은 오늘날의 현대 분류에도 적용되어 여전히 살아 있으나 많은 전문가들은 순수하게 완전한 계층은 이상적으로나 가능한 것임을 인식하고 있다. 그럼에도 불구하고 지식 도메인들은 계층 구조를 선호하며 지식 표현의 이론적 배경으로 삼고 있다.

계층 구조는 다음의 엄격한 구조적 요건을 필요로 한다:

• 포괄성: 최상위 클래스는 가장 포괄적인 클래스로 분류의 도메인을 기술하는 것이며 모든 중위 클래스와 하위 클래스를 포함한다.

• 종/속(Species/differentia): 정확한 계층은 상하위 클래스 간에 일반적으로 종/속관계의 한 가지 유형의 관계만을 갖는다.

• 상속: 해당 클래스에 포함된 개체에 참(true)인 모든 속성은 중하위 클래스에 포함된 개체들에게도 참으로 적용된다.

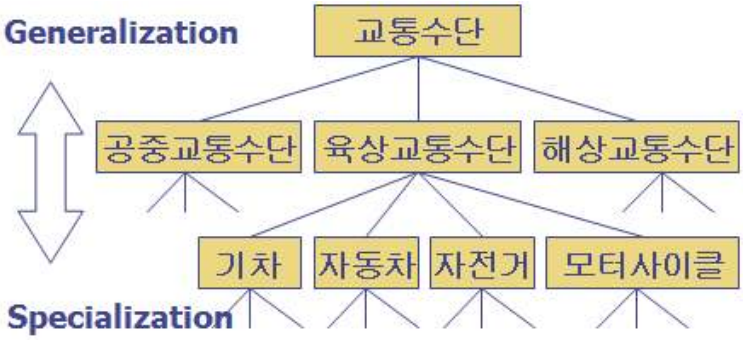
• 이행성(Transitivity): 속성이 상속되기 때문에 모든 하위 클래스들은 바로 위의 상위 클래스뿐만 아니라 그 위의 모든 상위 클래스의 구성요소가 된다.

• 연관 및 구별에 대한 체계적이고 예측 가능한 규칙: 하나의 클래스로 그룹화되는 규칙은 사전에 결정되어진다. 마찬가지로 하위 클래스의 구별을 위해 생성되는 규칙 또한 이미 결정되어진 방식으로 이루어지므로 예측 가능하다.

• 상호배타성: 하나의 개체는 오직 하나의 클래스에만 속한다.

• 필요충분조건: 하나의 클래스에 속하기 위해서 개체는 규정된 필요 속성을 가지고 있어야 하며 그 다음 충분한 근거를 구성하고 나면 그 개체는 해당 클래스에 반드시 속하게 된다.

이와 같은 특성을 지니고 있은 계층 분류는 다음의 몇 가지 이유로 지식의 표현과 발견에 있어 지속적인 관심을 받고 있다.



<그림 3-1> 계층 구조 분류의 사례

• 완전하고 포괄적인 정보: 계층 분류는 집합 및 구별을 위한 모든 법칙이 연역적으로 이루어지기 때문에 대개는 매우 포괄적인 분류 특성을 갖는다. 즉, 사전에 구조가 설정되고 고안자는 많은 양의 개체의 범위, 속성 그리고 서로 유사하거나 차이가 중요한 기준들에 대해 알고 있어야 한다.

• 기호의 상속과 경제성(Inheritance and economy of notation): 계층의 형식주의는 많은 복합적인 속성의 경제적 표현을 가능하게 한다. 각 속성은 매 수준마다 반복될 필요가 없다.

• 추론: 속성이 상속되기 때문에 충분하지 않은 정보만으로도 추리가 가능하다. 예를 들어 다른 동물들과의 관찰과 비교를 통해 고양이가 포유류임을 가늠할 수 있다면 암컷인 경우 새끼를 낳아 젖을 먹인다는 것을 추론할 수 있다.

• 실질적 정의: 계층 분류는 다른 유형보다 정의 표현이 우수한 것으로 평가된다. 개체가 무언가와 어떻게 유사한지 또는 어떤 중요한 측면에서 차이가 있는지 표현하는 방식을 제공함으로써 개체의 속성과 범위를 효과적으로 기술할 수 있기 때문이다.

• 상위 수준의 관점과 전체론적 시각: 분류 구조 설정이 기본적이고 의미 있는 구별을 나타낸다면 전체로서 분류 체계는 그것이 표현하고 있는 현상의 시각화를 제공한다. 전체론적인 관점은 연구자로 하여금 개별적인 사례로부터 단계적으로 더 넓은 문맥을 볼 수 있게 함으로써 때로 지식 생성을 촉발하기도 한다.

그러나 모든 지식 도메인이 계층에 의해 표현되는 것은 아니며, 다음과 같은 문제점을 가지고 있다:

• 복수의 계층: 현대의 관점에 의하면 세상은 더 이상 하나의 실체로 볼 수 없다. 즉 대부분의 현상은 표현의 문맥과 목적에 따라 여러 가지로 이해될 수 있는데 때

로 속성과 관계들이 중복되기고 하고 분리되기도 한다. 예를 들어 '개'는 '동물로서의 개'와 '애완동물로서의 개' 즉 사회적 관점과 동물학적인 관점으로 이해될 수 있다. 따라서 서로 얽혀있거나 복수 계층 구조의 상호 연결이 필요하다.

• 복수 및 다양한 기준: 많은 정보를 점점 복잡해지는 구조 안에 포함시키는 것은 실제적인 제한이 있으며, 계층 구조는 두 개의 매우 다른 기준을 구별하도록 설계하기 어렵다.

• 완전하고 포괄적인 지식의 결여: 계층은 구조 전반에 걸쳐 모든 개체 및 개체들 간의 관계를 보여주려고 하기 때문에 미리 도메인의 완전한 지식을 요구한다. 그러나 적절한 근거나 증거 없이 계층 구조에 밀어 넣은 경우 또는 쇠퇴하거나 새롭게 출현하는 분야에 대해서는 이들을 적절하게 조정하여 계층구조화하기 어렵다. 결과적으로 포괄성이 결여되고 분류 결과가 비논리적이 된다.

• 계층 수준의 차이(Differences of scale): 이행성과 상속의 원칙을 유지하기 위해서는 하나의 계층에 있는 모든 개체는 동일한 개념적 수준에 있어야 한다. 그러나 같은 주제어라 하더라도 각기 다른 수준의 정의를 포함할 수 있으며 그와 같은 차이를 하나의 분류 내에서 수용하기는 쉽지 않다.

• 이행성의 결여(Lack of transitivity): 계층은 속성이 구조를 따라 아래로 내려가는 것을 요구한다. 만약 A가 B의 하위 클래스이고 B가 C의 하위 클래스이면 A가 B의 하위 클래스여야 한다. 그러나 인간을 둘러싼 현상을 인식하는 방법이 항상 이와 같이 분명하게 되지 않는다. 이와 같은 현상은 추론을 어렵게 할 수 있다.

• 클래스 포함 규칙의 지나친 엄격성: 개체들이 항상 필요-충분 조건을 따르는 것은 아니다. 순수한 계층 구조에서 개체들은 반드시 분명하게 클래스에 속해야 하지만 실제 개체들은 둘 이상의 클래스에도 속할 수도 있다. 나아가 하나의 클래스에 속한 개체들은 서로 공통되는 속성을 일부 공유할 수 있으나 모두가 동일한 속성을 공유하는 것은 아니다.

3.1.2 트리 구조(trees)

개체와 그들의 관계를 표현하는데 사용되는 또 다른 유형의 분류 유형은 트리 구조이다. 트리 구조는 구별을 위해 계층 구조에서와 같은 특정 규칙에 기반하여 클래스를 구분하지만 상속의 규칙을 따르지는 않기 때문에 개체들 간의 종속 관계가 성립되지 않을 수도 있다. 트리 구조에서 개체들은 군대의 계급과 같이 위계를 가지고 있으나 반드시 클래스 간의 속성을 공유하는 것은 아니다. 지식표현의 면에서

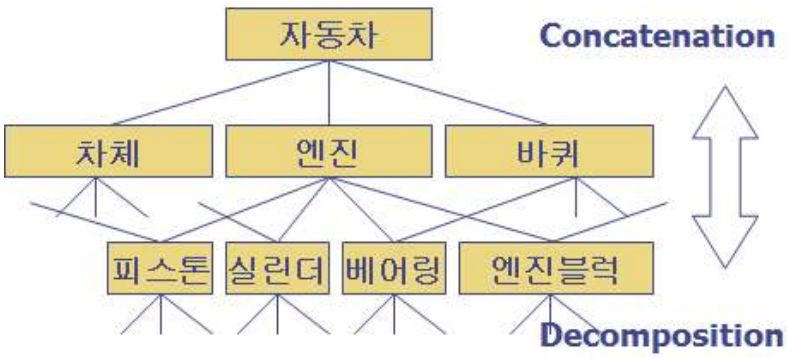
트리구조는 특정한 관계와 관련된 개체의 비교 분포를 잘 보여줄 수 있고, 무엇이 가장 위에 있는지 무엇이 가장 아래 있는지 명령의 연결고리를 보여준다. 일부 특권과 책임에 관한 추론도 가능하지만 실용적인 지식에 근거를 두고 있기 때문에 추론성이 약하다.

트리 구조에서의 개체의 관계 유형은 전체-부분 관계에 의해 연관되는 경우이다. 즉 각각의 클래스는 구성요소들로 나뉘고 또 다른 하위 구성 요소로 나뉘게 된다. 사실 이와 같은 관계는 계층구조의 전체-부분 관계와 차이가 없기 때문에 둘 다 '계층 구조'라고 하기도 한다. 두 가지 모두 보다 일반적인 것으로부터 특수한 것으로 내려가는 기호를 전달하고 있지만, 정확한 속성에 의한 추론을 이끌어내기 위해서는 표현의 구축에 대한 주의가 필요하다. 트리 구조는 다음과 같은 형식적인 요건을 필요로 한다:

• 완전하고 포괄적인 정보: 계층 구조와 같이 트리 구조에서도 개체들은 미리 결정되어진 구조에 포함된다. 개체에 대한 지식은 분류의 범위와 중요한 구별 기준을 결정하기 위해서 비교적 완전하게 갖춰져야 한다.

• 구별을 위한 체계적이고 예측 가능한 규칙: 트리의 일반적 구조는 개체들 간의 관계에 의해 결정된다. 특정 도메인의 지식 표현에 따라 전체/부분 관계, 원인/효과, 시작/결과, 과정/생산 등에서 적절한 관계가 선정될 수 있다.

• 인용 순서(Citation order): 계층 구조와 트리 구조 모두 구별 규칙을 적용할 때 순서를 결정하는 것이 중요하다. 이러한 결정들이 분류의 표현적인 호소력을 갖추게 되므로 '시작점'이 가장 중요하다. 만약 시작점이 적절하지 않으면 트리 구조의 나머지 역시 어색하고 지식을 제대로 반영하지 못하게 된다.



<그림 3-2> 트리 구조 분류의 사례

트리구조는 다음과 같은 이유로 유용한 지식의 표현방법이 된다:

• 하이라이트/디스플레이 관계: 이것은 트리 구조의 주된 장점이며, 클래스의 패턴에 따라 개체간의 관계를 한정하거나 중요하게 강조하여 볼 수 있도록 제시해준다.

• 거리: 트리 구조는 개체 간의 거리(물리적, 은유적 거리)를 나타내준다.

• 개체의 상대적 빈도: 이 특징은 계층 구조에도 적용되는 것으로 하나의 클래스 아래에 많은 수의 개체가 군집을 이루는 경우, 그들 간의 구별을 위한 새로운 규칙의 생성이나 발견에 대한 기회가 많아진다는 것이다. 반대로 하나의 범주 아래 개체 수가 매우 적게 되면 다른 범주로 통합될 필요가 있거나 구분 논리를 다시 생각할 필요가 있다.

트리 구조의 문제점은 계층 구조의 문제점과 유사하며 다음과 같다:

• 경직성(Rigidity): 트리 구조는 일반적인 나무의 형태로 사전에 개체간의 관계와 인용 순서가 결정되어지기 때문에 새로운 개체들이 추가되기 어려운 구조이다.

• 정보의 일방적 흐름: 전체-부분 관계가 표현되기는 하지만 정보가 수직적으로 위에서 아래의 방향으로 흐른다. 종/속 관계 규칙이 적용되지 않으며, 특히 복수 방향의 복합 관계를 표현하는데 적절하지 않다.

• 선택적 관점: 계층을 가지고 있는 트리 구조는 특정 관계만 강조하여 보여줄 수 있으며 다른 흥미로운 관계를 보여주지 못한다.

3.1.3 패러다임(paradigms)

개체들이 동시에 두 개 속성의 교차에 의해 기술되어지는 구조를 말한다. 결과로 나타나는 매트릭스(또는 패러다임)는 속성이 교차될 때 개체의 특성이 존재하는지의 여부를 보여 준다. 패러다임의 형식적 요건은 다음과 같다:

• 두 방향 계층 관계: x축과 y축의 두 개 방향에 의해 각 셀의 개체들이 관련된다. 열을 따라서 개체들은 하위 클래스의 다른 개체들과 관련되어지며, 행에 있는 개체는 동일한 하위 클래스에 있는 다른 개체와 연관된다. 그러나 포괄적인 관계에 있는 개체들은 속성을 서로 상속하지 않는다.

• 축은 두 개 관심사의 속성 표현: 각각의 축은 하나의 속성을 표현하고 있으며 두 개의 축은 한 번에 두 개 차원으로 개체들이 분류되는 것을 보여준다.

• 셀이 비어있거나 둘이상의 개체를 가질 수도 있다: 두 개 속성이 교차되는 것

을 보여줄 뿐만 아니라 개체들의 존재상태(존재, 부재, 복수)를 보여준다.

다음은 패러다임의 제한점이다:

• 도메인의 지식을 요구: 패러다임의 표현력은 기본 개념을 반영하는 두 개축에 표현되는 속성의 적절한 선택에 달려 있다. 이론이나 모형에 의한 범위를 사용하는 패러다임은 합의된 틀에 의한 기술(description)에 기반하기 때문에 해당 도메인의 지식을 반영하는 좋은 구조이다.

• 제한된 관점: 범위가 잘 선택되면 타당한 설명을 지니게 되지만 또한 무엇을 보이게 할지 범위를 제한하는 필터가 될 수도 있다. 개체의 문화적 정의가 달라지고 다른 관점이 제시된다면 다른 분석적인 결과를 만들어 낼 것이다.

• 제한된 설명력: 패러다임이 쌍으로만 범위를 적용하기 때문에 현상의 완전한 그림을 생성하기 어렵다. 패러다임이 잠재적으로 수직, 수평적 계층 관계의 풍부한 표현을 사용하는 반면 복합적인 설명이 어렵다.

3.1.4 팻싯 분석(faceted analysis)

패싯식 분류는 실제 다른 유형의 표현적인 구조라기보다 분류 과정에 대한 접근법이라고 할 수 있다. 패싯의 개념은 세상을 보는 관점이 둘 이상이라는 믿음에 근거하며, 분류는 신축적이며 새로운 현상을 수용할 수 있어야 한다고 본다.

패싯식 분류는 랑가나단의 저작에 뿌리를 두고 있는데 그는 어떤 복합적인 개체도 수많은 관점 또는 패싯으로 살펴볼 수 있다고 하였다. 랑가나단은 다섯 가지 기본 범주로 개성(Personality), 물질(Matter), 에너지(Energy), 공간(Space), 시간(Time)을 제시하였다. 수십 년간 랑가나단의 패싯은 많은 맥락에서 재해석되어져 왔으나 놀랍도록 시간의 테스트를 잘 견뎌내었다. 결과적으로 패싯은 서로 다른 컴퓨터 소프트웨어, 특허, 도서, 예술 작품의 객체들을 분류하는데 사용되어 왔다.

모든 패싯식 분류가 랑가나단의 규정된 기본 범주를 사용하는 것은 아니지만 그들이 공통적으로 지닌 것은 분석의 과정이다. 패싯식 접근은 다음의 단계를 따른다:

• 패싯 선택: 기술(description)을 위한 중요한 기준을 미리 결정한다. 이것들은 패싯 또는 기본 범주를 형성한다.

• 패싯 전개: 각각의 패싯은 자신의 논리와 근거 그리고 분류 구조를 이용하여 전개되고 확장될 수 있다. 예를 들어 물질 패싯은 계층 구조에, 공간 패싯은 트리 구조에 적용될 수도 있다.

• 패싯을 이용한 개체 분석: 개체를 분석할 때 적절한 패싯으로부터 디스크립터(용어)를 선택하고 문자열을 형성시킨다. 분석은 개체를 계층 구조에서와 같이 구별화된 범주로 세분하는 것이 아니라 모든 각도에서 객체를 살펴보는 과정임을 주목해야 한다.

• 열거 순서 전개: 분류된 객체를 조직할 때 주요 속성이 되는 첫 번째 패싯을 선택하고 다른 패싯과의 열거 순서를 결정해야 해야 한다.

패싯식 접근은 세상을 어떻게 조직되어야 할 것인지에 대한 현대의 요구에 매우 적절한 것으로 평가되며 특히 다음과 같은 이유로 유용한 도구가 된다:

• 완전한 지식을 요구하지 않는다: 패싯식 체계를 구축할 때 체계에 수용되는 개체 및 패싯 간의 완전한 범위를 알 필요가 없다. 특히 새롭게 출현하는 분야 또는 변화하는 분야에 유용하다.

• 폭넓은 수용성(Hospitable): 새로운 개체를 용이하게 수용할 수 있다.

• 신축성(Flexibility): 패싯식 체계는 수많은 독립적인 속성에 의해 각각의 객체를 설명하기 때문에 이러한 속성들은 끊임없이 신축적인 방법으로 적용될 수 있다. 이러한 신축성은 새롭고 흥미로운 결합으로 이어질 수도 있으며 후조합 방식의 검색을 지원한다.

• 표현력(Expressiveness): 패싯식 접근은 패싯이 어휘와 구조를 통합하는데 자유롭기 때문에 패싯으로 표현되는 지식에 가장 적합하다.

• 강력한 이론을 요구하지 않음: 패싯식 분류는 전체적인 구조를 가지고 있지 않기 때문에 이론적 연결고리의 근거를 필요로 하지 않는다. 기본 범주가 잘 기능할 수 있는 한 그 때 그 때 구축될 수 있다.

• 이론적 구조 및 모델의 다양성 수용

• 복수 관점: 패싯식 접근의 가장 유용한 특징은 다양한 관점에서 개체를 볼 수 있게 한다는 것이다. 이 특징은 계층 구조와 트리 구조에서 결여되어 있는 부분이다.

패싯식 분류의 제한점은 다음과 같다:

• 적절한 패싯 구축의 어려움: 개체를 신축성 있게 추가할 수 있는 반면 도메인이나 잠재적 이용자에 대한 지식 없이 기본 패싯을 설정한다는 것은 매우 어려운 일이다.

• 패싯 간의 관계 부족: 이론화 및 모델 구축의 관점에서, 패싯식 분류는 다면적인 기술에 유용하지만 의미있는 방법으로 다양한 패싯을 연결하는 임무를 잘 수행

하지 못한다.

• 시각화의 어려움: 계층조나 트리조 특히 패러다임은 시각적으로 개체들과 그들 간의 관계를 분명한 방식으로 보여줄 수 있다. 반면 패싯 분류는 실제 기술적인 내용에 훨씬 복합적인 표현이 결합되어 있음에도 불구하고 동시에 한두 개 정보만 보여 줄 수 있는 구조를 가지고 있다.

3.2 구성방식에 의한 유형

3.2.1 연역적-귀납적

자료를 분류할 때 적용하는 시점을 기준으로 연역적 분류와 귀납적 분류로 구분할 수 있다. 연역적 분류는 하나의 전제로부터 개별적인 명제들을 이끌어 내는 방식이다. 즉 일반적인 것에서 특수한 것으로 세분하는, 환언하면 주목하는 성질의 판단기준을 미리 결정한 상태에서 분류하는 방식으로 선천적 또는 이론적 분류로도 지칭되고 있다. 반면에 귀납적 분류는 개별적인 사례로부터 하나의 명제를 도출해 내는 방식이다. 즉 특수한 것에서 일반적인 것으로 종합하는 다시 말해서 대상자료를 관찰하면서 어느 성질에 주목할 것인지를 판단하여 분류하는 방식으로서 후천적 또는 경험적 분류학고도 한다. 실제 분류행위에서는 양자의 복잡한 상호관계를 전제로 이루어진다.

3.2.2 하강적-상승적

하강적 분류는 어떤 자료군을 성질이 상이한 몇 개의 영역으로 구분한 다음, 다시 각 영역에 속하는 자료 중에서 이질적인 여러 개로 세분하는 이른바 이질성에 주목하여 행하는 분류이며 구분이라고도 한다. 가령 문학작품을 시, 희곡, 소설, 수필 등으로 구분하고 소설을 다시 애정소설, 과학소설, 탐정소설, 역사소설 등으로 분류하는 것을 말한다.

반면 상승적 분류는 동일한 성질을 지니는 자료를 몇 개의 그룹으로 군집하고, 각 그룹이 지니는 상위의 동질성에 따라 다시 대별하는 소위 동질성에 주목하는 분류행위이다. 예컨대, 한국과 미국의 시, 소설, 수필을 각각 고대 문학과 현대 문학으로 분류하고 다시 한국 문학과 미국 문학으로 대별하는 경우이다.

3.2.3 자연적-인위적

자연적 분류는 자연 현상을 객관적 속성 또는 연관성에 따라 분류하듯이 피분류체의 성질을 분류 기준으로 삼는 것을 말한다. 가령 동물을 척추의 유무에 따라 척추 동물과 무척추 동물로 구분하고 어패류에 속하는 고래를 어류가 아닌 포유류에 귀속시키는 것을 자연적 분류라 한다. 반면 인위적 분류는 피분류체인 자료를 자의적 또는 임의적 기준에 따라 분류하는 것이다. 예컨대 자료의 종류를 일반 도서, 연속간행물, 학위 논문, 시청각자료, 데이터베이스로, 장정에 따라 고서와 현대서로, 크기를 근거로 문고본, 국판 또는 크라운판, 대형본으로 전자출판물을 오프라인 자료와 온라인 자료 등으로 구분하는 것은 일종의 인위적 분류에 해당한다.

3.2.4 서지-서가

서가 분류와 서지 분류는 분류의 목적이나 용도를 기준으로 구분하는 유형이다. 동서양을 막론하고 모든 도서관계와 문헌정보학계가 자료 분류를 유형화할 때 가장 주목하는 방식이다. 그 배경은 대상 자료의 주제를 분석하고 다수의 패싯 기호를 조합하는 기본적인 목적이 서지 작성과 서가 배열 등에 있기 때문이다.

서지 분류는 카드 목록이나 책자 목록을 편성(편집)하기 위한 목록상의 분류이기 때문에 복수의 주제를 취급한 자료인 경우, 모든 주제를 분류 기호로 표현할 수 있다. 복수의 주제명이나 분류 기호로 검색할 수 있어 주제 검색을 지향하는 분류로 간주되지만, 일단 서지류에 등재되면 이동이 어렵다는 측면에서 고정식 배열법으로 지칭되고 있다. 19세기 중반 무료공개 및 공비 운영을 원칙으로 하는 근대 도서관이 형성되면서 도입된 개가제에 편승하여 서가 분류가 도입되기 전까지, 즉 고대에서 19세기 중반까지를 서지 분류 시대로 지칭하고 있다.

서가 분류는 주제나 형식에 따라 자료를 서가상에 배열하기 위한 상관식 배가법의 방식을 취하고 있다. 분류로서 동일한 주제를 기술한 자료가 동일한 서가에 군집 배치되고 유사한 주제자료가 인접 배치되므로 이용자의 서가 접근을 통한 브라우징 및 검색 기능이 제고되며, 자료의 주제별 입수정도 및 구성 내용을 쉽게 파악할 수 있다. 제약점으로는 자료라는 물리적 단위를 대상으로 하기 때문에 복수의 주제를 취급한 자료인 경우라도 하나의 특정 주제로 분류된다. 즉, 주제의 수와 무관하게 항상 하나의 특정 주제 아래 분류되고 제2주제 이하는 분류의 대상에서 제외된다.

서가 분류는 19세기 후반부터 이용자가 직접 서가에 접근하는 환경, 즉 개가제가 도입됨에 따라 등장하였고 이후에 대중화되었다. 1870년 해리스(W.T. Harris)가

창안한 분류법은 서가 분류와 서지 분류의 일원화를 시도한 최초의 분류법으로 회자되고 있으며, 1876년에 발간된 듀이(M. Dewey)의 DDC에 의해 서지 분류와 서가 분류가 일원화되었다. 따라서 오늘날 대부분의 도서관계가 채택하고 있는 자료분류법(DDC, UDC, LCC, CC, KDC, NDC 등)은 서가 분류와 서지 분류에 동시에 적용되는 분류법이다.

4. 자료 분류표의 유형 및 선정

4.1 기호법에 의한 유형

4.2 표시 방식에 의한 유형

4.3 대상 범위에 따른 유형

4.4 자료 분류표의 선정

4. 자료 분류표의 유형 및 선정

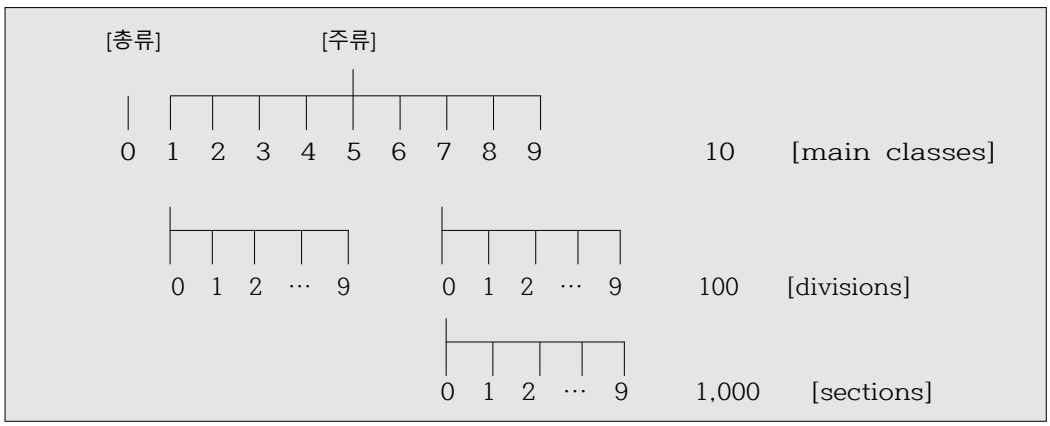
4.1 기호법에 의한 유형

4.1.1 십진분류표

십진분류표는 순수 아라비아 숫자를 사용하여 주제의 내용을 10구분씩 점진적으로 세분한 것을 말한다. 즉, 모든 지식을 9개(1~9)의 주류로 구분하고, 기타를 0에 배정한 다음에 각각의 대그룹(주류)을 다시 10개의 하위 그룹(강목)으로, 각각의 하위 그룹을 다시 10개의 소그룹(요목)으로 세분한 것을 말한다(<그림 4-1> 참조>. 현재 십진분류표는 세계적으로 범용되고 있으며 대표하는 분류표로는 DDC, UDC, KDC, NDC 등이 있다.

일반적으로 십진분류표는 기호 구성이 단순하고 신축적이며, 조기성이 풍부하고 실용적이다. 또한 분류 기호로 주제와 개념의 상하 관계를 파악할 수 있고, 숫자를 기호로 변환하였기 때문에 이해와 기억이 용이하며 국제적인 적용성을 지닌다. 그 외에도 대개 상관 색인을 제공하고 있어 사용하기 편리하고 동일 분류표를 사용하는 도서관 간의 상호 협력이 용이하다.

반면 십진분류표는 총류를 제외하면, 실제 구분지가 9개로 한정되기 때문에 형식적이며, 지식 전체가 기계적으로 구분된다. 또한 분류체계가 기호법에 예속되기 때문에 분류 기호가 길어질 수밖에 없다. 그리고 동일한 계위의 숫자가 이미 분류표에 사용된 경우에는 새로운 주제의 삽입이 어려워 주제 배열에서 논리성이 떨어질 수 있다는 점과 비십진분류표에 비교할 때 전개력이 약한 단점을 지니고 있다.

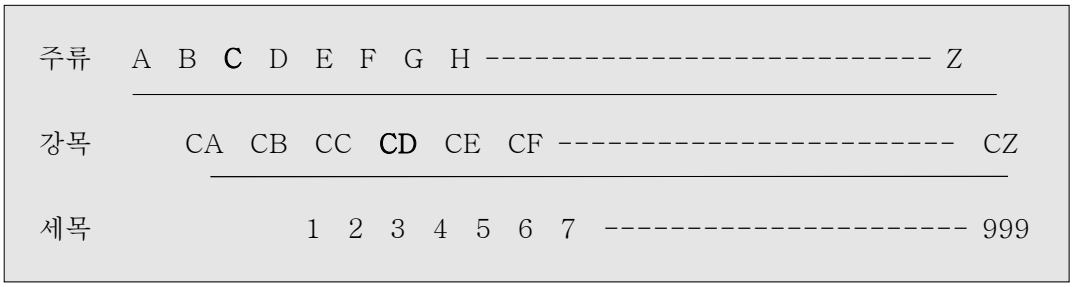


<그림 4-1> 십진분류표의 기본구조

4.1.2 비십진분류표

비십진분류표는 숫자 1~9까지를 사용하는 십진식을 제외한 모든 분류표를 총칭한다. 대개 알파벳 문자만 사용하는 것과 문자에 숫자와 기호를 조합하는 것이 있으며 LCC, CC, EC, SC, BC 등이 대표적이다.

비십진분류표는 대규모 도서관이나 주제지향적 전문도서관에서 많이 채택한다. 가장 큰 이유는 십진식 분류표보다 전개력이 뛰어나 상세한 분류가 가능하기 때문이다. 예컨대 LCC는 알파벳 문자 26개에 의한 주류와 강목의 산술적 전개 항목이 최대 676(26×26)인 반면에 DDC는 100(10×10)에 불과하다. 비십진식 분류표의 기본구조는 <그림 4-2>와 같다.



<그림 4-2> 비십진분류표의 기본구조

비십진식분류표의 장점은 기호에 구애받지 않고 분류 체계를 합리적으로 구성할 수 있으며, 전개력이 우수하다는 것이다. 또한 분류 기호의 계층 숫자가 십진식보다 간단하고 동시에 하위 구분지의 숫자도 조정이 가능하다. 전개 폭이 넓어 미래의 새로운 주제를 삽입시키기 위한 공기호의 여분이 많아 효율적이다. 그러나 기호법이 복잡하여 청구 기호 라벨에 기재하기 어렵고 서가 배열을 위한 시간이 많이 소요되며, 조기성이 부족하여 기억하기 어려운 단점이 있다.

4.2 표시 방식에 의한 유형

4.2.1 열거식 분류표

열거식 분류표는 지식 전체를 소수의 학문 영역으로 범주화하고 각 영역에서는 주제 간의 관계 유형에 따라 최고의 유개념에서 마지막 종개념까지 계층적으로 전

개한 분류표로서 모든 개념을 단일 체계로 조직한 것이다. 전통적인 열거식 분류표는 일련의 상호배타적인 유(주제)를 열거하는 계층 구조의 원리를 따르며 연역적, 하향식 방식의 분류표이다.

열거형 분류표의 장점은 기호 시스템이 복잡하지 않고 비교적 단순하여 분류표를 구조화하기 쉽기 때문에 오랫동안 도서관 커뮤니티에서 널리 사용되어 온 매우 친숙한 분류 도구라는 점이다. 또한 예상되는 주제 간의 관계를 미리 분류표에 제시하는 경우 추정되는 (복합)주제에 대한 적합한 기호를 찾기가 용이하다.

그러나 열거형 분류표는 개념간의 다양한 관계를 정확하게 제시하기 어렵고 주제의 특수한 관점을 충분히 표현하지 못하며 개념간의 유연한 결합이 불가능하여 유용성의 한계를 지니고 있다. 열거형 분류표의 단점으로는 다음과 같은 사항이 지적되고 있다.

• 주제의 전개 범위가 제한되어 있어(특히 십진분류에서) 비논리적이며 인위적이어서 결과적으로 개념의 계층 구조를 표현하는데 무리가 있다.

• 모든 주제나 지식을 완전하게 제시할 수 없고 오류와 간격이 필연적으로 따르게 된다. 따라서 용어의 변화와 지식의 발전을 수용하기 위해서는 지속적인 개정작업이 필요하다.

• 상이한 주류(학문 영역)간의 결합이 불가능하며 모든 복합 주제를 미리 제시해야 하는 한계를 지니고 있다. 아울러 복수의 주제를 취급한 문헌의 경우 특정 주제와 관련된 학문 영역을 먼저 선정해야 한다.

열거식 분류표의 예로는 LCC를 비롯하여 DDC, UDC, KDC, NDC 등이 있으며, 위에서 지적된 바와 같이 어떤 분류표도 모든 학문과 지식을 완벽하게 계층 구조로 열거할 수 없기 때문에 오늘날 대다수 열거식 분류표는 몇 가지 보조표를 구비하여 이를 부분적으로 합성시키는 패싯 분석에 의한 조합 방식을 접목하고 있다.

4.2.2 분석합성식 분류표

분석합성식 분류표는 어떤 특성을 근거로 지식 또는 주제의 전분야를 기본 주제로 세분하여 본표에 최소 분류 항목만 표시하되 나머지는 합성하도록 구성한 분류표를 말한다. 이 유형의 분류표는 열거식 분류표가 복수 주제, 복합 주제, 합성 주제를 기호화하지 못하는 한계를 극복하려는 노력의 결과로 등장하였다. 이를 대표하는 사례가 패싯 분류표이며 1933년 랑가나단이 창안한 콜론분류표(CC)가 그 전형이며, 주제 간의 계층관계나 지식의 분화과정은 물론, 관련 주제 간의 결합력이

우수하여 분류학에 크게 기여한 분류표로 평가되고 있다.

분석합성식 분류표는 기본주제 및 패싯을 분석하고 해당 기호를 합성하는 패싯식 구조 원리를 따르고 있으며 귀납적, 상향식 방식의 분류표이다. 분석 합성식 분류에서는 개념을 일정한 특성에 따라 구성요소로 분석하는데 이를 패싯 분석이라고 한다. 패싯은 속성, 차원, 행위 등으로 조직화된 범주이며 분석합성식 분류에서 개념은 상이한 패싯 아래 독립적으로 제시되고 필요에 따라 합성된다. 즉, 개념간의 계층 구조를 일일이 제시하는 대신, 문헌의 주제를 소수의 관점(특성이나 측면)으로 분석하여 파악하고 이들 관점마다 분류 기호를 부여한 다음 이를 일정한 원칙에 따라 합성하여 주제를 표현하는 분류 과정을 따른다.

일반적으로 분석합성식 분류표의 장점은 열거식 분류표보다 본표의 분량이 적고, 특수 주제나 합성 주제를 모두 나열할 필요가 없기 때문에 편찬하기 쉬우며, 기존의 주제나 개념에 신주제를 조합할 수 있다는 점이다. 그러나 본표가 간단함에도 불구하고 다양한 패싯 기호와 조합방식을 채택함으로써 기호가 길어지고 복잡하며, 열거 순서의 결정 문제가 분류자의 심적 부담을 가중시키고 실용성을 약화시키는 단점을 가지고 있다.

4.3 대상 범위에 따른 유형

4.3.1 종합 분류표

종합(일반) 분류표는 지식의 전분야를 망라적으로 체계화한 분류표를 말한다. 현대의 주요 분류표인 LCC, DDC, UDC, KDC, NDC, BC, CC 등은 채택한 기호법, 표시방식, 구조가 어떠하든 종합 분류표에 해당한다. 대부분 각 주제의 구성은 어느 한 분야에 편중되거나 치우치지 않고 똑같이 평준화되어 전개되어 있다. 따라서 모든 학문 영역이나 지식 분야에서 출간된 다양한 자료를 수집하는 경우에는 종합분류표를 사용해야 하므로 대다수 대학 및 공공도서관에서 많이 채택하고 있다.

대부분의 도서관에서 공통적으로 사용되도록 육성할 목적으로 분류표에 필요한 모든 조건을 구비하여 간행되는 표준 분류표는 종합(일반) 분류표 중에서 널리 채택되어 사용되어진 분류표를 근간으로 한다.

4.3.2 특수 분류표

특수 분류표는 특정 주제(의학, 농학, 법학, 음악, 경영 등)나 특수 자료(지도, 음

반, 신문기사 등)에 적용할 목적으로 개발된 분류표를 말한다. 특수 분류표는 자료가 1) 특정 주제 분야에 한정된 경우, 2) 특수 형태의 자료가 한정된 경우, 3) 특정 문서 자료에 한정된 경우에 해당된다. 주로 전문 도서관이 채택하며, 대학 도서관도 종합대학 내의 단과대학이나 특수 목적 대학에 적용된다.

4.4 자료 분류표의 선정

4.4.1 실물 자료를 위한 분류표

도서관이나 정보센터에서 다루는 일차적인 정보자원은 특정한 형태를 지닌 실물 자료로서 대표적으로 인쇄 자료(단행본자료, 연속간행물)와 비도서자료(비디오녹화자료, 음악자료, 지도, 마이크로 자료 등)가 해당된다. 이들 자료는 전통적으로 자료분류표에 의해 분류되고 서가에 배열되어 보존·관리되며 이용되는 특성을 지니고 있다. 실물 자료를 위한 보편적인 선정 기준을 제시하면 다음과 같다(윤희윤 2013, 61);

• 도서관의 집서 규모가 적고, 사서 및 이용자의 편리성이 중요할 경우 KDC와 같은 단순한 분류표나 DDC, UDC 등의 간략판을 채택하는 것이 바람직하다.

• 기존의 장서 구성이나 수집하는 자료의 주제가 광범위할 경우에는 종합 분류표를, 한정된 주제나 특정 주제에 치중하는 경우에는 특수 분류표를 사용하는 것이 합리적이다.

• 종합 분류표를 채용할 경우, 서양서 비중이 높은 대학 도서관은 DDC를 학문 영역이 매우 제한적인 단과대학 중심의 도서관은 LCC를, 국내서 비중이 90%를 넘는 대다수 공공 도서관은 KDC를 선정하는 것이 무난하다.

• 특수 또는 전문도서관 가운데 과학기술분야 비도서 자료가 압도적으로 많은 도서관은 UDC를 특정 주제 도서관(자동차연구소자료실, 섬유도서관 등)은 LCC를, 기타 특수 목적의 도서관(신문사자료실, 특허자료실, 의료원자료실 등)은 특수 분류표를 채택하면 무리가 없다.

4.4.2 디지털 자원을 위한 분류표

디지털 시대에는 모든 도서관이 인쇄 자료, 비도서 자료와 전자출판물을 동시에 수용 · 제공하는 하이브리드형 자료 공간으로서의 정체성을 유지해야 한다. 일반적으로 도서관에서 정보 검색 서비스를 위해 메타데이터를 작성하는 대상의 전자 자

원의 유형으로는 전자 저널, Web-DB, E-book 등이 대표적이며, 도서관에 따라 인쇄 자료(고서, 학위 논문 등)를 아카이빙하는 디지털 자료 등이 포함된다. 이밖에도 인터넷을 통해 제공되는 원문 형태의 수많은 웹 자원이 전자 자원을 구성하고 있으며 구글, 야후, 네이버 등과 같은 인터넷 포털 서비스 기관을 통해 자원의 탐색이 이루어지는 경우가 많다.

전자 자원의 수용과 더불어 이들을 분류할 것인가 어느 정도 범위까지 분류가 가능한가 분류하게 된다면 어떠한 분류표를 적용할 것인가와 같은 문제에 직면하게 된다. 자료 분류의 일차적 기능이 실물 자료의 체계적인 배가를 통한 이용자 접근 검색을 지원하는 수단으로 한정한다면 전자 자원을 분류할 필요가 없다. 그러나 모든 관련 주제를 군집하여 통합 검색을 지원하는 도구로 간주한다면 전자 자원도 분류할 필요성이 있다. 다만 저널 기사(articles), 원문(full-text)과 같은 방대한 양의 정보를 모두 분류할 수 없기 때문에 텍스트의 자동 범주화에 대한 논의는 구별하여 검토할 필요가 있다.

먼저 도서관을 통해 전자 자원의 분류 도구 선정과 개발의 방법으로는 기존의 인쇄 자료용 분류표를 적용하는 방안, 그것을 수정 · 사용하는 방안, 그리고 별도의 웹 자원용 분류표를 개발하는 방안이 있다. 분류표의 개발에는 전문 지식과 추론 능력, 주제에 대한 이해와 더불어 많은 시간과 노력이 수반되며 이런 점에서 기존의 자료 분류표를 이용하여 전자 자원을 분류할 경우 다음과 같은 효과를 기대할 수 있다(Koch 1997);

• 브라우징: 분류표를 통해 주제의 구조와 용어에 친숙하지 않은 이용자에게 브라우징 기회를 제공한다. 아울러 온라인 환경에서 항해 보조 수단으로 분류표를 사용할 수 있다.

• 탐색 범위의 확장과 축소: 분류표는 계층적으로 조직되어 있기 때문에 그 구조 내에서 광의 또는 협의 주제를 선택하여 탐색의 범위를 조정할 수 있다.

• 맥락적 내용: 분류표의 사용으로 탐색의 내용을 알 수 있다. 예컨대, 동형이의어의 문제를 부분적으로 해결해 준다.

• 다국어적 접근 허용의 가능성: 분류표는 특정 언어와는 독립된 기호(숫자나 알파벳 문자)를 사용하기 때문에 동일 자원에 복수의 언어로 접근할 수 있다. 특정 언어의 탐색어를 입력하면 변환 언어와 같이 분류표의 언어(분류 기호)로 변환되고, 해당 주제에 관해 특정 언어로 된 자원을 검색할 수 있다.

• 데이터베이스의 분할 및 처리: 대규모 분류표는 필요에 따라 하위 요소로 논리적으로 분할 될 수 있다.

• 상호운용성: 현존하는 많은 분류표는 기계가독형으로 이용할 수 있어 인쇄 자료뿐만 아니라 웹 자원에 적용할 수 있으므로 상호운용성을 보장한다.

• 안정성: 대부분의 분류표는 주기적인 개정을 통해 최신성을 유지하고 있어 분류도구로서의 안정성을 확보하고 있다.

• 친숙성: 이용자는 인쇄자료 분류표에 대해 인지하고 이해하고 있기 때문에 웹 자원에 적용하더라도 친숙하게 이용할 수 있다.

자체 웹 자원용 분류표의 개발은 주로 웹에서 브라우징 검색을 위한 넷 디렉토리용 도구로써 사용된다. 넷 디렉토리에서는 브라우징을 위해 주제를 계층 관계에 따라 제시한 분류표를 사용하는데 그 예로 Yahoo에서는 웹 자원에 계층적으로 접근할 수 있는 자체 분류표(온톨로지)를 개발하여 사용하고 있다. 넷 디렉토리를 이용한 탐색은 편리하고 원하는 자료에 바로 접근할 수 있다는 장점이 있다. 그러나 넷 디렉토리에 의한 탐색은 수작업 분류에 기초하고 있기 때문에 접근할 수 있는 웹 자원이 제한되어 있다는 한계를 가지고 있다(김태수 2000, 241-243).

반면 브라우징 검색과 더불어 웹 자원의 탐색을 위한 키워드 기법은 탐색 엔진을 도구로 사용한다. 탐색 엔진은 전 주제영역을 대상으로 대량의 웹 자원을 구문과 키워드로 탐색하는데 특히 검색의 포괄성과 검색 능력의 우수성을 장점으로 한다. 그러나 모든 주제와 모든 이용자를 동일한 수준에 두고 설계된 탐색 엔진의 대부분은 탐색 결과가 없거나 너무 많은 결과를 제시하게 함으로써 부적합 문헌이 상당수 출력되는 한계를 지닌다. 따라서 전문 주제 영역보다는 주로 일반적인 관심 영역에 관한 정보를 탐색하는 데에 유용한 것으로 평가되고 있다.

이에 따라 전 주제영역을 탐색 대상으로 하는 탐색 엔진과는 달리 소규모이면서 탐색 능력이 우수한 특정 주제 영역의 탐색 서비스가 개발되고 있다. 이 서비스에서는 주제 전문가를 통해 질적으로 우수한 자원을 선정, 기술하고 주제를 부여한다. 동시에 시소러스나 분류표와 같은 공식적인 지식 구조를 통해 선정된 자원에 주제로 접근할 수 있는 서비스를 제공하고 있다(김태수 2000, 244-245).

5. 자료 분류표의 요건과 개발 단계

5.1 자료 분류표의 기본 요건

5.2 자료 분류표의 개발 단계

5. 자료 분류표의 요건과 개발 단계

5.1 자료 분류표의 기본 요건

분류표가 그 본래의 목적을 원활하게 달성하기 위하여 갖추어야 하는 구성상의 기본 요건은 다음과 같다(김포옥, 백항기 2011, 41-42)

• 분류하고자 하는 대상이나 목적, 방법 면에서는 전적으로 학문 분류를 따르기는 어려우나 지식을 세분화하고 생성하고, 전개시켜 체계화하기 위해 학문 분류를 참고해야 한다.

• 분류원리는 한 가지로 일관성을 유지시켜야 하며, 군집된 부분 집합 내의 상호 배타성을 지켜야 하며 피분류체의 망라적 범위, 구분의 점진적 전개가 조성되어야 한다.

• 모든 자료의 주제를 분류할 수 있는 망라성, 학문의 세분화에 따르는 신축성, 그리고 특별 항목의 세분 전개를 위한 보완성을 갖추는 동시에 구체적이고 정교해야 한다.

• 분류 체계의 계층 구조 내지 하위 단계에 열거되는 분류 명사는 용어상으로나 의미상으로 명확하게 구분되어야 한다. 명사 앞뒤에 관련된 유사어나 동의어가 명시되어 있으면 분류에 혼동을 주기 쉽다. 동시에 시대 변화에 따르는 학문상의 통일된 용어를 사용해야 한다.

• 본표 외에 각종 조기성 보조표가 구비되어야 한다. 주제가 다수인 자료, 형식이 주제보다 우선되는 자료, 관점이 두 개 이상 취급된 자료 등을 해결하기 위한 지침과 보조 기호 등의 용례법을 제시하여야 한다.

• 단순하고 간결한 분류 기호를 마련해야 한다.

• 분류표의 이해를 위한 설명과 사용 지침서를 가능한 상세히 마련하여 이용 편의를 도모할 수 있어야 한다.

• 모든 분류표는 색인이 준비되어 있어야 한다.

• 분류표는 영구적인 기관에 의해 항상 개선, 유지되고 관리되어야 한다.

5.2 자료 분류표의 개발 단계(김태수 2000, 130-144)

5.2.1 개념 단계의 원리

(1) 구분의 특성

분류에서는 지식을 소수의 학문 영역으로 나누고 이 과정을 필요한 단계까지 연속적으로 반복하게 된다. 따라서 구분 단계마다 적용되는 특성을 선정해야 하고 특성의 적용 순서에 따라 조직하는 것이 분류의 핵심 과정이다.

일반적인 구분 원리를 제시하면 1) 학문 영역간의 경계가 분명해야 한다. 2) 분류표에서 일단 선정된 범주는 영속성을 지녀야 한다. 일시적인 필요나 상황에 따라 범주를 구성하는 것이 아니라 영속적인 특성을 적용해야 한다. 3) 지식의 구분 특성은 실증적이고 유용해야 한다. 4) 구분의 특성은 분류의 목적과 일치해야 한다.

(2) 특성계열

분류에서 지식을 구분하기 위해서는 단일 특성에 의한 단일 범주화만이 아니라 연속적인 범주화가 필요하다. 분류에서 주류 즉, 범주는 상호 배타적이어야 하고 그러기 위해서는 한 번에 하나의 특성만을 적용하여 하위 범주로 전개해야 한다. 만약 두 개의 특성을 동시에 적용해야 하는 경우 분류의 목적에 더 중요하고 우선 적용할 것인가를 결정해야 한다. 특성의 순서를 규정하는 데 적용되는 절대적인 원칙은 있을 수 없으며 대부분 복수의 특성이 적용된다. 특성의 순서는 관련 학문 영역의 합의와 대다수 이용자들의 접근 방식을 반영한 결과이다. 문학 분류에서 언어-문학 형식-시대의 열거 순서를 고려할 수 있다. 특성의 순서를 결정하고 나면 분류표에서 일관성을 확보하기 위해 이 순서를 일관되게 적용하는 것이 중요하다.

(3) 동위류의 조직

하나의 특성을 적용하여 개념을 구분하게 되면 일련의 대등한 수준의 개념 즉, 동위 개념을 얻을 수 있다. 동위류는 지식 구조에 포함되는 동위 수준의 모든 개념을 포용할 수 있어야 한다. 그러나 지식의 역동적인 성격으로 인해 특정 수준의 구성요소 전부를 완전하게 제시하는 것은 사실상 불가능하다. 따라서 분류표에서는 새로운 개념이 출현할 때 이를 수용할 수 있는 유연성이 요구된다. 일단 동위류를 구성하게 되면 분류표에서는 이를 직선 형식으로 배열하게 되는데 배열 순서는 분류 목적에 유용해야 한다. 가장 중요한 관점이 무엇인가를 고려하는 것이다.

(4) 하위류의 조직

분류표의 주제나 구분지의 계층 관계에 적용되는 규범이다. 외연 감소 규범은 개념의 외연은 각 구분 단계마다 감소되고 외연이 증가할수록 내포는 감소되는 것을 의미한다. 상위 주제에서 하위 주제로 갈수록 개체의 수는 감소하고 특성은 증가됨

을 알 수 있다. 반면 조정 규범은 유사한 주제나 개념이 상이한 계열에 출현할 때는 항상 대등하게 취급되어야 하는 규범이다. 일관된 순서를 유지하기 위해서는 특정 주제의 계열을 구성할 때 공통 구분지와 같이 동일한 표를 사용하고 유용한 순서를 적용하여 계열마다 대등한 순서를 유지하게 하는 것이다.

분류에서 내포는 분류의 대상인 범주에 적용한 연속적인 특성의 수에 따라 표현된다. 하위류를 특정한 순서로 배열하기 위해서는 항상 기본 원칙을 동일하게 적용해야 한다. 단일 계층에 속하는 동위류의 배열 기준은 분류표를 설계하는 사람의 자유지만 상이한 수준에서는 항상 왼쪽에서 오른쪽으로, 그리고 위에서 아래로 연결되어야 한다. 이것은 특정 류(class)는 그 동위류보다는 하위류와 더 밀접한 관계를 가진다는 분류 규칙을 따른 것이다.

6. 자료 분류법 사례

6.1 듀이 십진 분류법

6.2 미국의회도서관 분류법

6.3 콜론 분류법

6.4 전자 정보원 분류와 자료 분류법

6. 자료 분류법 사례[1)]

6.1 듀이 십진 분류법

6.1.1 개요

DDC(Dewey Decimal Classification)로 불리는 듀이 십진 분류법은 창안자인 듀이와 십진식 분류라는 데서 붙여진 명칭으로 1876년 44페이지의 초판이 발행된 이후 약 7년의 갱신주기에 따라 2011년 현재 총 4권의 4,276여 페이지에 달하는 23판이 발행되었다.

2002년에 발행된 22판의 특징은 웹 환경에 그 내용을 소개한 첫 번째 판으로서 국제적으로 다양한 이용자들의 요구를 충족시키고 상호 협력을 증진시키는 것과 분류에서의 효율성과 정확성을 높이려는 것이다. 2011년의 23판은 인쇄 형태의 발행과 함께 전자 버전인 WebDewey를 동시에 사용할 수 있도록 하였다. 특히 WebDewey는 지난 수년간 DDC를 이용하거나 관심 있는 세계 각국의 문헌 분류표 이용자들에게 DDC의 갱신된 내용을 전달해주는 주요 수단으로 활용되어 왔다. 또한 DDC 기호와 미국 국회도서관 주제명 표목표(LCSH), 미국 의학도서관 주제명 표목표(MeSH), BISAC 표목들과 매핑이 가능하도록 하였다. 이처럼 DDC 제23판은 전세계 이용자들과 상호 작용에 의해 다수의 새로운 분류 기호와 토픽들을 갱신하였으며, 분류 실무자의 능률성을 제고하기 위한 도구가 된다는 견해와 함께 기계적 표현과 적용요건을 충족시키는 문헌 분류표가 되고 있다(<표 6-1> 참조).

1) 대다수 자료분류법은 기원 후에 등장하였다고 할 수 있다. 서양의 중세 시대에는 수도원 도서관이 자료를 생산하고 보존하는 주체였다. 문예 부흥기에는 근대 도서 분류법의 시조인 동시에 최초의 서지학적 목록이며 백과전서의 효시인 '세계 서지(Bibliotheca Universalis)'가 게스너(Conrad von Gesner, 1516~1565))에 의해 1548년에 완성되었다. 1605년에는 영국의 사상가 베이컨(Francis Bacon, 1561~1626)이 학문을 체계화하고 지식을 기억(사학), 상상(시학), 이성(이학)으로 분류하였으며 도서관의 초창기 분류에 영향을 미쳤다. 18세기 초에 와서는 이성적이고 논리적인 분류대신 실용적인 배열 체계가 중시되기 시작했고 19세기에 이르러 분류학자들은 종래의 주관적이고 이상적인 구분방식 대신 지식분류의 객관적인 기준을 설정하기 위해 합리적인 접근방식을 적용하였다. 대표적인 예로 해리스(W.T. Harris, 1835~1909)의 분류법, 커터(C.A. Cutter, 1837~1903)의 전개 분류법(EC: Expansive Classification), 브리스(H.E. Bliss, 1870~1955)의 서지 분류법(BC: Bibliographic Classification) 등이 개발되었다. 이들 분류법은 20세기 이후 현대 분류법의 토대가 되었으며 1876년에는 듀이(Melvil Dewey, 1851~1931)가 DDC(Dewey Decimal Classification)를, 1901년에는 미국 의회도서관이 LCC(Library of Congress Classification)를, 1905년에는 국제서지학회가 DDC 5판을 저본으로 UDC(Universal Decimal Classification)를, 1933년에는 랑가나단(S.R. Ranganathan, 1892~1972)이 CC(Colon Classification)를 발간하였다. 동양의 대표적인 분류법으로는 중국의 사분법(經史子集), 일본의 일본 십진 분류법(NDC: Nippon Decimal Classification) 그리고 한국의 한국 십진 분류법(KDC: Korean Decimal Classification) 등을 들 수 있다.

<표 6-1> DDC 제23판과 제22판의 특징 비교

23판	제22판
• 새로운 토픽들과 특정 분야의 중요한 갱신 • 전세계 이용자와의 상호 소통을 통한 분류 기호 공지 • 사람의 집단(groups of people)에 대한 표현 점검 • 표준 세구분표의 상당부분의 개정 • 이중 표목과 불균형한 범위 조정	• 웹 환경에 대한 적용 • 지속적인 갱신 • 새로운 분류기호 및 토픽의 추가 • 국제적 시각의 확대 및 상호협력 • 분류담당자의 능률성 도모

DDC는 오늘날 138개 이상의 국가에서 사용되고 있으며, 20만개관 이상이 사용하는 국제 표준 분류표이며, 30개국 이상의 언어로 번역되었다. DDC를 분류 도구로 사용하는 국가 중에는 60개국 이상이 국가 서지를 작성할 때 DDC를 적용하고 있다(<표 6-2> 참조). 특히 1988년에 OCLC가 DDC의 저작권을 소유하면서 문헌 분류의 효율성과 정확성을 제고하기 위해 전통 있는 문헌 분류 시스템으로서의 친밀성과 일관성을 유지하여 왔다. 미국의 경우 공공 및 학교 도서관의 95%, 대학 도서관의 25%, 전문 도서관의 20%가 분류 도구로 DDC를 사용하고 있고, 영국은 대학 도서관의 85%, 공공 도서관의 99%가 사용하고 있다. 한국의 경우 2011년 현재 대학 도서관의 63.5%가 사용하는 가운데 공공 도서관과 전문 도서관에서는 극히 일부만 사용하고 있다.

<표 6-2> 국가서지 작성에 DDC를 적용하는 국가 현황

대륙	국가
아메리카	캐나다, 브라질, 페루, 멕시코, 콜롬비아, 칠레, 베네수엘라, 볼리비아, 피지, 바베이도스, 버뮤다, 가이아나, 트리니다드 토바고
유 럽	영국, 독일, 프랑스, 오스트리아, 이탈리아, 스위스, 그리스, 아일랜드, 아이슬랜드, 노르웨이, 터키, 몰타
아프리카	남아공화국, 가나, 나이지리아, 잠비아, 케냐, 탄자니아, 자메이카, 짐바브, 리비아, 보츠와나, 감비아, 나미비아, 시에라리온, 스와질란드, 자이레
아 시 아	인도, 태국, 인도네시아, 이란, 이라크, 필리핀, 말레이시아, 파키스탄, 베트남, 라오스, 방글라데시, 네팔, 시리아, 스리랑카, 카타르, 팔레스타인
오세아니아	호주, 뉴질랜드, 뉴기니

6.1.2 포맷과 기본 구조

4권으로 되어 있는 DDC 제23판 인쇄본의 주요 구성 부분은 다음과 같다.

- 제1권: 서언 및 서문, 제 23판의 새로운 특징, 서론, 용어 해설, 서론과 용어 해설의 색인, 매뉴얼, 보조표, 제 22판과 제23판의 비교표
- 제2권: DDC 개요표(summaries), 본표(000-599)
- 제3권: 본표(600-999)
- 제4권: 상관 색인

6.1.3 주류의 배열 원리

DDC의 주류 구성의 토대는 베이컨의 분류 체계와 이를 도치시켜 창안한 해리스의 역베이컨식 분류 체계에서 기원한다. 베이컨은 학문 전체를 인간의 지식과 신학으로 대별한 다음에 전자를 다시 정신 능력에 따라 기억, 상상, 이성으로 구분하고, 각각에 대응하는 학문을 역사(history), 시학(poesy), 철학(philosophy)으로 규정하였다. 그러나 해리스는 베이컨의 학문 분류 체계를 도치시켜 과학(철학 포함), 예술, 역사의 순으로 배열하되, 헤겔의 세구분을 적극 수용하였으며, 이를 듀이가 답습하였다.

듀이는 해리스의 주류 구성을 더 구체화하여 모든 학문 분야를 9개 영역(100-900)으로 구분하고, 기타 귀속성이 불분명한 학문(서지학, 문헌정보학, 박물관학, 언론학 등)과 토픽(일반백과사전, 연속간행물, 각종 단체 등)을 총류(000)에 배정하였다.

DDC가 총류, 철학, 종교를 제외한 학문분야를 사회 과학-자연 과학-응용 과학-인문 과학의 순으로 배치한 기본 구조는 다른 많은 분류표(UDC, KDC, NDC 등)의 주류 배열체 계에 영향을 미쳤다.

6.1.4 주요 특징

① **학문에 의한 분류**: 주제보다는 학문에 의한 분류를 채택한 학문적 분류표이며 관점 분류법(aspect classification)을 지향함으로써 여러 자료가 동일한 주제를 기술하더라도 접근 또는 관점이 다르면 각각의 관점을 중시하여 분류할 수 있어 선택의 폭이 넓고 관점에 따른 군집력이 강하다. 예를 들어 '가족(Families)'이라는 하나의 주제는 법률이나 레크리에이션, 사회 복지 등을 포함한 여러 학

문 분야에서 이를 다루게 된다. 즉 가족의 '법률적 문제'는 법률학의 일부로서 346.015에 해당하며, '레크리에이션'은 오락의 일부로서 790.191에 해당하며, '사회·복지 문제'는 사회복지학의 일부로서 362.82에 해당한다.

② **계층적 구조**: DDC는 계층적 분류표(hierarchical classification)로서 학문이나 주제의 관계를 나타내기 위해서, 일반적인 것들로부터 시작하여 점차 구체적인 것들로 전개된다. 그러나 이것은 하나의 원칙으로 채택된 것으로서, 분류표에 일반적으로 적용된다는 의미이며, 반드시 모든 경우에 완전하게 적용된다는 것은 아니다.

③ **십진식에 의한 전개**: DDC에서는 모든 지식을 주류하는 열 개의 광범위한 학문 분야로 구분하고, 이들을 계속적으로 강, 목, 세목의 단계로 십진식으로 세분한다. 이론상으로 보면 계속적인 십진식 전개에 의해 무한히 전개할 수 있다. 십진식 전개는 모든 주제를 항상 열 개씩 세분해야 한다는 점에서 근본적으로 불합리성을 내재하고 있다. 그러나 숫자만 사용하는 순수 기호법을 채택하고 있다는 점과 더불어, 십진식 전개는 그 편리성 때문에 DDC가 국제적으로 채택될 수 있게 한다.

④ **조기성의 도입**: 분류 기호의 조기성(mnemonics)은 "분류 체계가 개념을 표현할 때 어떤 개념이 출현하는 위치에 관계없이 이를 동일한 기호로 표현하여 기억을 돕는 것, 또는 그 반대로 분류 체계에서 동일 기호는 동일한 개념을 표현하도록 하여 기억을 돕는 것"이다. 조기성이 부여된 분류 기호는 분류 담당자가 문헌에 대한 분류 기호를 부여할 때 기억을 도와주며, 분류표와 색인을 참고해야 하는 일을 상당 부분 줄여준다. 또한 분류표의 길이를 줄여 주며, 유한한 배열 구조에는 일관된 순서를 택할 수 있도록 해준다.

⑤ **본표 및 보조표에 있는 요약(summaries)**: 요약은 내부 목차에 해당하며, 각 분류 항목의 구성 체계 및 배열 구조 전체를 개관한 것으로 다른 분류표에 없는 매우 편리한 접근수단이다.

6.1.5 분류 기호 합성 방식

분류 기호를 구성할 때 <표 6-3>과 같이 다양한 유형의 합성 방식을 제공한다.

<표 6-3> DDC 분류 기호 합성 방식과 용례

	분류기호 합성방식	기호합성(예)
	본표의 전주제 기호 합성	• 예술도서관: 026(기본기호) + 700(본표의 예술) = 026.7
②	본표의 다른 강목 또는 세목의 일부기호 합성	• Plant physiology: 571.2 + 212(582.12) = 571.212
③	본표의 동일한 강목 또는 세목의 일부기호 부가	• 도서관 희귀서 수집: 025.28 + 16(025.341-025.349 중025.3416이 희귀서) = 025.2816
④	내부표(internal table) 합성	• 여성을 위한 고용서비스: 362.83 + 84(내부표 362-363 아래의 84 고용서비스) = 362.8384
⑤	패싯기호(facet indicator) 합성	• Beneficial insects: 595.7 + 1(중심표목 592-599 아래의 지시에 따라) +63(591.63) = 595.7163
⑥	보조표(T.1~T.6) 합성	• General statistics of India: 31(기본기호) + 54(T.2 India) = 315.4

6.1.6 분류 규칙

① 일반규칙: 열거식 분류표의 태생적 한계인 합성 주제를 모두 기호화하는데 따른 분류 오류와 비군집화를 방지하기 위하여 다음과 같은 분류 규칙을 제공하고 있다.

- 적용 규칙: 어떤 자료가 2개 주제(대상, 요소)의 상관관계(영향, 인과)를 기술한 경우에 영향을 받은 주제 또는 결과에 해당하는 주제에 분류해야 한다는 규칙으로서, 다른 어떤 규칙보다 우선 적용된다. 예를 들어 '타고르가 한국 시문학에 미친 영향'을 다루고 있는 문헌은 한국시에 분류한다.
- 선행 규칙(first-of-two rule): 어떤 자료에서 두 주제를 동등하게 다루고 있고, 서로에 대한 소개나 설명이 이루어져 있지 않을 경우에는 해당 자료를 그 분류 기호가 DDC 본표에서 첫 번째로 나타나는 주제에 분류한다. 예를 들어 일본과 러시아를 동등하게 다루고 있고 일본이 먼저 논의되고 서명에도 첫 번째로 기재되어 있는 경우 947 러시아가 952 일본보다 앞에 오기 때문에 러시아 역사(947)에 분류한다.
- 삼자 규칙(rule of three): 모두가 동일한 상위 주제의 세목에 해당하는 셋 이상의 주제를 다루고 있는 문헌은 어느 한 주제가 다른 주제들보다 더욱 완전하게 다루어지지 않는 한, 이 주제들을 모두 포함하는 첫 번째 상위 기호에 분류한다. 예를 들어 터키 역사(956.1), 이라크 역사(956.7) 및 시리아 역사(956.91)를 다루고 있는 문헌은 이 세주제의 상위에 해당하는 중동사(956)에 분류한다.

- 0의 규칙(rule of zero): 어떤 자료를 분류한 결과, 복수의 분류기호가 모두 적합한 것으로 생각될 경우에는 '0'을 수반하는 세구분과 '1-9'로 시작하는 세구분 가운데 후자를 우선적으로 선택한다.

② 열거 순서와 우선순위: 복수 주제의 측면이나 특색을 분류 기호에서 처리하거나, 그 가운데 어느 하나를 선택해야 하는 경우 일관성 유지에 유용한 열거 순서(citation order)와 우선순위(preference order)를 제공하고 있다. 열거 순서는 분류 기호를 합성할 때 어떤 주제나 유에 나타나는 여러 패싯이나 특성들을 어떤 순서로 결합할 것인가를 결정하게 해주는 순서이다. DDC에서 일반적으로 채택하고 있는 열거 순서는 학문 → 주제와 각 계층의 세분 주제 → 지리 및 시대 세분 → 표현 형식의 순으로 되어 있다. 다만 000 총류의 경우는 학문 → 표현 형식 → 언어나 장소의 순서를 채택하고 있고, 800 문학에서는 학문 → 언어 → 문학 형식 → 시대의 순서를 채택하고 있다.

제19판부터 도입된 우선순위는 어떤 주제의 여러 특성을 기호의 합성을 통해서는 충분히 나타낼 수 없는 경우에, 그와 같은 특성을 나타내는 기호 가운데 어느 기호를 선택해야 할지 지시해 주는 순서이다.

③ 임의 규정(선택 조항): 분류에서는 어떤 주제를 둘 이상의 방식으로 다룰 수도 있다. 이러한 경우에 어떤 주제를 공식적으로 채택된 방식과는 다른 방식으로 배열하고자 하는 도서관에 융통성을 부여하기 위해, 특정 주제에 대해 임의기호(optional numbers)를 제공하고 있다.

6.2 미국 의회도서관 분류법

6.2.1 개요

LCC(Library of Congress Classification)으로 불리는 미국의회도서관분류법은 1901년 개요가 만들어진 이후 2009년 기준으로 모든 주제를 21개로 대별하여 49권 이상의 분책 형태로 발행되는 세계 최대의 열거식 분류표인 동시에 일반분류표이다. 이들 각 분야는 서로 다른 전문가들에 의해 어떤 하나의 주류나 하위류 또는 하위류의 일부분을 독립해서 출판하고 있다. 따라서 LCC는 전체로는 일반분류표이지만 각각의 주류 측면에서 볼 경우 서로 조정하여 작성된 일종의 특수분류표이기도 하다.

LCC는 본래 의회도서관 소장자료를 분류할 목적으로 개발된 시스템이었으나 비

십진분류표로서 전개능력이 우수하기 때문에 현재는 많은 미국의 대규모 대학도서관, 연구도서관 등에서 사용되고 있는 표준분류표로 간주되고 있다. 북미 도서관 가운데 소장책수가 50만권을 상회하는 경우로 한정하면 62% 이상이 LCC를 적용하고 있으며, 영국에서는 26개관 정도가 사용하고 있다. 동양권에서는 이란의 국가도서관과 장서규모가 큰 다수 대학도서관에서 채택되고 있다.

이처럼 국내외에서 LCC를 분류도구로 채택하는 주된 이유는 하위류나 세목의 전개 능력이 DDC보다 우수하고, 분책형 분류표가 주제도서관(분관)을 위한 분류도구로 적합하며, 수시로 개정·보완되기 때문이다. 특히 도서관 외부관계자들이 이용할 수 있는 MARC 목록데이터를 통해 LCC 번호를 쉽게 확인할 수 있다.

6.2.2 본표의 기본 구조

LCC의 모든 본표는 다음과 같이 총 7개 부분으로 구성되어 있다.

- 서문(preface): 각 주류표의 역사적 배경과 주제 범위를 설명
- 개괄표(broad outline): 각각의 주류표 아래에 전개된 하위류(subclass), 예를 들어 L(Education) 아래의 LA · LB · LC 등과 같이 2~3개 문자의 리스트를 제시
- 개요표(detailed outline): 하위류 아래의 세목, 예를 들어 LA 5-25, LA 31-133 등의 리스트 제시
- 주류표(the schedule): 본표에 해당하는 각 주류표의 전체적인 내용을 열거
- 보조표(auxiliary tables): 주류 전체가 아닌 개별 주류에 선택적으로 적용되는 5종의 보조표(form, geographic, chronological, topical subdivision, combination)
- 색인(index): 대부분의 주류표가 알파벳 색인을 제공하고 있으나, 일부의 주류 및 하위류는 색인이 없기 때문에 종합색인이 존재하지 않음
- 부록(supplementary pages): 각 주류표의 마지막에 개정사항을 상세하게 기술

6.2.3 주류의 배열 원리

LCC의 주류 배열은 학문 분류의 수용보다는 LC의 장서 구성에 따른 편리 위주와 도서관의 실용성을 우선적으로 고려한 것으로 볼 수 있다. 그러나 LCC의 주류 배열은 전적으로 임의에 의한 것이라기보다 어느 정도 전개 분류법(EC: Expansive

Classification)의 주류 체계를 참고한 것이며, 과학, 기술 분야에 비해서 상대적으로 사회 과학 특히 역사 분야가 많은 비중을 차지하고 있다.

주류의 구성 체계는 모든 지식 분야가 주요 학문에 대응하는 주제로 A에서 Z까지 구분되어 있다. 신학문을 삽입하거나 주류의 신설을 대비하여 5개 알파벳 대문자(I, O, W, X, Y)를 공기호로 남겨 두었으나, 국립의학도서관 분류법(NLMC)이 LCC의 의학 주류표인 R(General medicine) 대신에 미배정 기호(Qs-QZ, W)를 사용하여 세분하였기 때문에 사실상 공기호는 4개이다. 주류 배열의 개요는 <표 6-4>와 같으며 이를 특성별로 군집하면 <표 6-5>와 같다.

<표 6> LCC의 주류배열

A	General Works	M	Music
B	Philosophy, Psychology, Religion	N	Fine Arts
C	Auxiliary Sciences of History	P	Language and Literature
D	World history and history of Europe, Asia, Africa, Australia, etc.	Q	Science
E-F	History of the Americas	R	Medicine
G	Geography, Anthropology, Recreation	S	Agriculture
H	Social Sciences	T	Techonology
J	Political Sciences	U	Military Science
K	Law	V	Naval Science
L	Education	Z	Bibliography, Library Science, Information resources(General)

<표 6-5> LCC 주류배치의 성격별 군집화

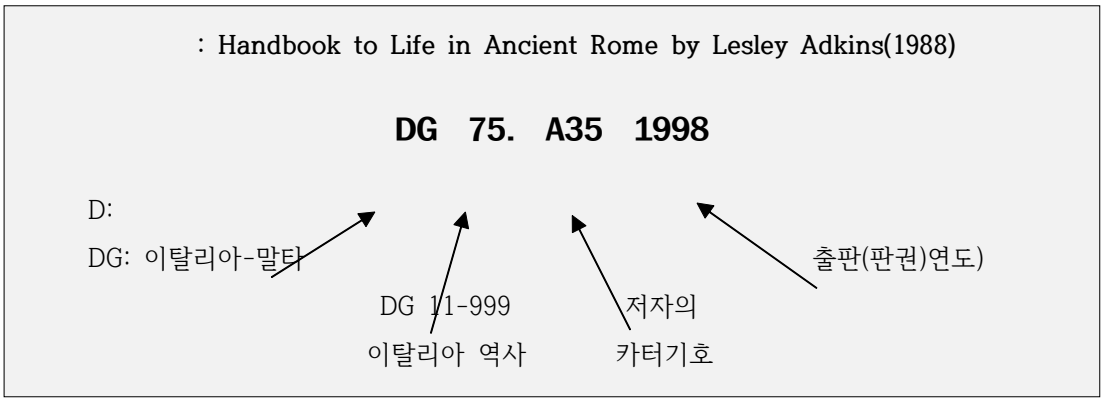
의미	영역의 성격
A ( )	특정 주제에 한정되지 않는 분야
B (철학, 심리학, 종교)	우주에 관한 인간의 이론과 정신
C-G (역사보조학, 역사, 지리, 인류학, 오락)	인간의 사회생활, 환경, 사고와 기록 등
H-L (사회과학, 정치학, 법률, 교육학)	인간의 경제적 및 사회적 발전
M-P (음악, 예술, 언어, 문학)	인간의 미적 및 예술적 활동
Q-V (과학, 의학, 농업, 기술, 군사학, 해군학)	이공계 영역
Z (서지, 문헌정보학, 정보자원)	문헌정보

다음으로 하위류의 경우, 주류 E-F와 Z를 제외하면 각각에 귀속되는 학문이나 주제가 느슨한 계층 구조로 세분되어 있으며 특히 P(언어와 문학), Q(과학), T(기술)에는 많은 하위류가 설정되어 있다. 그러나 하위류의 구성 항목 중에서 예컨대

QA(수학)는 많은 페이지를 할애한 반면에 LH(대학 및 학교의 잡지와 신문)는 매우 적게 배정하고 있어 LC가 소장한 장서의 구성 내용과 실용성을 우선적으로 고려하고 있음을 알 수 있다. 고 알파벳 대문자 2-3자로 하위류를 세분하고 있다.

6.2.4 주요 특징

① **기호법**: LCC는 4종류로 구성되는 혼합 기호법(mixed notation)을 채택하고 있다. 즉, 주류마다 1개의 알파벳 대문자를 적용하고 있으며 그 하위류에 2-3개의 알파벳 대문자(E, F, Z에는 적용되지 않음), 그리고 세목 이하에는 1-9999의 정수, 그리고 개별 자료를 구분하기 위하여 1개의 대문자와 숫자로 구성된 커터 기호(cutter number) 및 출판 연도로 구성되어 있다(<그림 6-1> 참조).



<그림 6-1> LCC의 혼합 기호법 구성 용례

② **비계층적 구조**: LCC의 중요한 특징은 일관된 계층 구조를 취하지 않고 있다는 점이다. 즉 세목의 전개에 있어서 십진식 분류표의 표준 구분, 시대 구분, 지역 구분 등의 보조표에 해당하는 내용과 본표의 내용도 상하 관계의 구분 없이 동일한 수준으로 정수 1-9999 중에서 선정하였다. 따라서 계층 구조와 그 관계를 기호법에 있어서는 무시할 수 있기 때문에 새로운 주제에 대응하기가 매우 용이하다.

③ **문헌적 타당성(literary warrant)**: 실용성에 초점을 두고 있는 LCC의 분류 원칙이 추상적인 지식을 조직하기 위한 개념에 바탕을 두는 것이 아니라 도서관에서 소장하고 있는 출판된 자료의 양에 따라 결정되는 문헌적 타당성에 근거하는 것이므로 다른 도서관의 장서에 일관성 있게 적용하는데 무리가 있을 수 있다. 그러나 실용적 측면에서 소장 자료의 이용을 극대화하기 위한 분류 체계를 작성하

기 위해서는 서비스하고자 하는 자료의 빈도를 조사하여 이를 근거로 작성하는 LCC의 원칙이 많은 의미를 지닌다.

6.3 콜론 분류법

6.3.1 개요

CC(Colon Classification)로 불리는 콜론 분류법은 1933년 인도의 랑가나단이 창안한 분석적 합성식의 대표적 분류법이다. CC는 열거식 분류표의 대표적인 DDC의 문제점 즉, 어떤 자료가 다양한 측면을 포함하고 있음에도 분류시 경우에 따라 주제의 특정 측면을 나타낼 수 없는 문제, 미래에 출현할 주제들이 더 많은 합성 개념을 가질 경우 이를 나타낼 수 없을 가능성이 높다는 문제, 모든 주제를 망라적으로 열거한다는 것이 불가능하다는 점을 극복하려는 노력에 의해 출현되었다.

따라서 콜론 분류법은 분류 대상물의 다면적 분석을 가능하게 하고 이렇게 다면적으로 분석된 요소들을 결합함으로써 자료가 지니고 있는 복합주제를 표현할 수 있게 한 분류표이다. 각 패싯에서 분석된 요소들을 결합할 때 콜론(:)을 많이 사용하여 콜론 분류표라고 하기도 하고 하나의 주제를 패싯으로 분석하여 패싯 분류표라고 부르기도 한다.

CC는 1933년 26개의 기본 주제로 구성되었던 초판이 발간된 이후 2판(1939년), 3판(1950년), 5판(1952년), 5판(1957년), 6판(1960년)의 개정이 있었다. 1987년 마지막 7판에 이르러 총 47개의 기본 주제가 776개로 세분됨으로써 학문의 분화와 증가를 잘 반영하고 있다. 본표는 공통세목과 특수세목, 공통 구분표, 지시 기호, 패싯 공식 등에서 많은 변화가 있었으며, 전판과 비교할 때 규칙이 간결하고, 공통 구분과 조기성 기호법의 다양성도 증가하였다. 그러나 인도에서는 제6판이 가장 범용되고 있다.

보급 현황을 살펴보면 CC는 다양한 패싯을 기호화하는 다면적 분류표이기 때문에 학술 연구의 대상으로 주목받아 온 반면, 도서관 현장에서는 분류 규정의 잦은 변경, 복잡한 기호 시스템, 난해한 용어로 인해 거의 채택하지 않고 있다. 1982~1983년의 조사결과, 인도의 CC 사용률은 27.5%(DDC 52%, UDC 12%, LCC 0.8%, 기타 7.7%)이며, 마드라스 대학도서관과 인도 노동성을 비롯한 약 2,500개관에서 사용하고 있다.

6.3.2 기본 구조

제7판의 제1권 본표의 구성은 다음과 같이 다섯 부분을 포함하고 있으며, F(색인), G(고전 분류표), H(고전 분류표의 색인)은 간행되지 않았다.

- A(서론: Introduction)
- B(초보자 가이드: Guidance to the Beginner)
- C(일반 규칙: General Rules)
- D(일반 세목 및 공통 세목: General divisions and common isolates)
- E(특수 세목: Special isolates)

6.3.3 주류 배열 원리

CC 주류(기본 주제)의 배열 원리는 프랑스의 암페르(A.M. Ampere, 1775~1836)의 학문 배열을 따랐다. 즉 암페르는 학문을 우주론과 정신 과학으로 양분하고, 기초 과학(물리학, 공학; 지리학, 광업; 식물학, 농학; 동물학, 축산학, 의학 등으로 세분), 실용 예술(useful arts), 응용 과학(applied science)의 순으로 배치하였다. 이 순서를 답습한 랑가나단은 기본 주제를 자연 과학, 인문 과학, 사회 과학의 순으로 배정하였으며, 제1권의 Part D에 설정된 구성내용 및 배열 체계는 <표 6-6>과 같다.

<표 6-6> CC의 기본 주제표

1	Communication Science	J	Agriculture
2	Library and Information Science	K	Zoology
3	Book Science	L	Medicine
4	Mass Communication	M	Useful arts
5	Exhibition technique	N	Fine arts
6	Museology/Museum technique	O	Literature
7	System Research, Systemology	P	Linguistics
8	Management Science	Q	Religion
A	Natural Science	R	Philosophy
B	Mathematics	S	Psychology
C	Physics	T	Education
D	Engineering	U	Geography
E	Chemistry	V	History
F	Chemical technology	W	Political Science
G	Biology	X	Economics(Macro-economics)
H	Geology	Y	Sociology
I	Botany	Z	Law

구체적으로 CC의 기본 주제는 추상적인 것에서 구체적인 것(예: 수학, 물리학, 생

물학, 의학의 순)으로, 자연적인 것에서 인위적인 것(예: 철학, 지리, 역사, 정치학, 경제학, 법률의 순)으로 배치하였다.

마지막으로 CC에도 다른 분류표(DDC, UDC, LCC, KDC 등)의 보조표에 해당하는 세목(구분지)이 있다. 특수 세목은 특정 주류에만 제한적으로 적용되며, 언어 구분기호, 시대 구분 기호, 지리 구분 기호와 같은 공통 세목은 일부 주류에 공통적으로 적용된다.

6.3.4 기호법

CC는 개정될 때마다 각종의 기호가 첨가 또는 변경되는 사례가 많다. 또한 알파벳 대문자 및 소문자는 물론 아라비아 숫자와 각종의 특수문자를 사용하여 혼합기호법을 채택하고 있다. 제7판의 기호법은 이전 판들에 비하여, 그리스 문자는 많이 삭제되었으나 기호의 종류와 구성은 보다 복잡한 양상을 띠고 있다.

기본 주제의 기호법은 23개의 알파벳 소문자(i, l, o를 제외한 a-z), 10개의 아라비아 숫자(0-9), 26개의 알파벳 대문자(A-Z), 원괄호로 묶은 숫자(( )), 지시기호(indicator digit), 붙임표(-), 별표(*) 등이 사용된다. 공통 세목과 특수 세목의 기호법으로는 23개의 알파벳 소문자, 아라비아 숫자, 26개의 알파벳 대문자 이외에 그리스 문자(Δ Delta), 3개의 전치기호(*, “ ←), 11개의 서수적 지시 기호(& ‘ . : ; , - = + ( ) →)가 사용된다.

이밖에도 불필요하게 중복된 열거를 피하고 분류 담당자들에게 자율권을 부여하기 위해 구분기호(device)를 적용하고 있는데 연대 기호표(chronological device), 지리 구분 기호표(geographical device), 주제 기호표(subject device), 열거순 기호표(enumeration device), 알파벳순 기호표(alphabetical device), 패싯 기호표(facet device), 상 기호표(phase device) 등 14개의 구분 기호표를 마련하고 있다.

6.3.5 패싯 구조

랑가나단은 처음으로 문헌정보학에 ‘패싯’이라는 용어를 도입하였으며 동시에 일관성 있게 패싯 분석 이론을 발달시킨 최초의 인물이다. 패싯 분석에 관련된 원칙들은 그가 만들어낸 원칙들 가운데 가장 강력하고도 영향력 있는 원칙들로, 20세기 주제 분석의 기초가 되고 있다는 평가를 받고 있다.

패싯은 간단하게 정의하면 범주(category)라고 할 수 있는데, 랑가나단은 패싯을 ‘일련의 특성에 바탕을 둔 기본류의 구분지들(divisions)의 총체’라고 하였다. 그리

고 일련의 특성에 의해 추출된 어떤 패싯이든 기본적으로 개성(Personality), 물질(Matter), 기능(Energy), 공간(Space), 시간(Time)의 5가지 범주에 포함되며 이를 적용하여 하위 주제로 분석할 수 있고 여러 측면의 하위 요소들은 연결 기호들에 의해 분류 기호로 작성된다(<표 6-7> 참조).

<표 6-7> CC 기본범주의 패싯 기호와 연결 기호

	의미		패싯기호	연결기호
Personality	: 본질적 속성	Who	[P]	,(comma)
Matter (property) (method) (material)	재료: 사물 : 특성 : 방법 : 재료	What	[M] [MP] [MM] [MMt]	;(semi-colon)
Energy	에너지: 활동, 작용, 공정 등	How	[E]	:(colon)
Space	공간: 지리구분	Where	[S]	.(dot)
Time	시간: 시대구분	When	[T]	'(apostrophe)

① 시간 범주[T]

주제 분석 과정에서 주제가 지닌 연대나 시대적 특성을 의미한다. CC에서는 특정 연대는 물론 겨울이나 주간, 야간과 같은 개념도 표현이 가능하다. 구체성 감소순으로 범주의 순서를 규정하고 있기 때문에 가장 추상적인 시간 범주가 가장 마지막에 위치한다.

② 공간 범주[S]

에너지 범주보다는 구체성이 약하지만 시간 범주보다는 더 구체성을 지닌 범주이다. 지표를 지형학적으로 또는 국가나 인구 밀도를 단위로 구분하는 것은 모두 공간 범주를 표현한 것이다.

③ 에너지 범주[E]

일반적으로 행위나 반응, 과정, 성질, 문제, 해결, 처리 등을 의미하는 범주이다. 언어학적으로 말하면 동사적 의미를 지닌 사항이 에너지 범주에 해당된다. 예컨대, 처리 과정이나 가르치고 치료하는 행위, 구조나 기능, 질병, 환경적 행위를 표현하기도 한다.

④ 재료 범주[M]

추진 방법이나 성질, 혹은 재료를 의미하는 범주이다. 에너지 패싯이 적극적인 것임에 비해 다소 수동적인 범주이다. 재료-성질[MP]는 개체가 지닌 기본적이고 고

유한 성질을 표현하는 패싯이며, 재료-방법([MM]은 각종 처리기법과 관련된 패싯이고, 재료-재료[MMt]는 재료나 매체를 의미하는 패싯이다.

⑤ 개체범주[P]

다섯 개의 기본 범주 중 가장 구체적이지만 확인하기 어려운 범주로서 마치 인간의 개성과 같이 다소 추상성을 지닌 범주이다. 주제를 구체적으로 제시하는 본질적 요소로서 이 개체 범주가 없으면 주제가 형성되지 않는다. 개체 범주는 주제 영역에 따라 사람이나 기관, 물질, 화합물, 동물이나 식물, 신체 기관, 언어, 종교 등이 될 수 있다.

6.3.6 패싯 공식

CC는 분석 합성식의 분류 원리를 채택하고 있으며 각각의 주류(기본 주제)를 몇 개의 기본 범주로 나누고 다시 각각의 범주 하에 세목을 배열하고 있다. 따라서 어떤 자료의 주제를 몇 개의 구성요소로 분석하고 각각의 패싯 내에서 구성요소들의 세목을 취사선택한 다음에 이들을 다시 합성하기 위해서는 일정한 결합 순서를 결정해야 한다. 지식을 구조화하는 주된 목적은 특정 자료의 검색에 있기 때문에 일관된 결합 순서의 유지가 중요하다.

이 기준을 패싯 공식(facet formular)이라 하며, 본질적으로 패싯은 그 주제 분야에서 차지하는 중요도나 개체 패싯의 여부에 따라 그 순서가 결정된다. 패싯의 순서는 일반적으로 구체성의 중가 원칙에 따라 결정되며, 제7판에 열거된 패싯 공식의 예를 들면 <표 6-8>과 같다.

<표 6-8> CC 기본 주제의 패싯 공식(예)

		패싯공식
2	Library and Information Science	2,[P];[MP]
G	Biology	E,[P];[MP]
J	Agriculture	J,[P],[P2];[MP]:[E]
O	Literature	O,[P],[P2],[P3],[P4]
T	Education	T,[P];[MP]

패싯 공식에서 첫 번째 PME의 조합을 제1회차라고 하며, 이들 범주는 다시 출현할 수 있다. 일반적으로 회차는 해당 패싯 앞의 숫자로 표현되고, 수준은 패싯 다음

의 숫자로 표현된다. 예를 들어 [2E]는 2회차 에너지를 의미하며 [P2]는 2차 수준의 개체 범주를 의미한다.

이상에 따라 실제 '1970년대 일본에서의 벼 바이러스 박멸'에 관한 주제를 분석하여 주제 요소를 기본 범주와 패싯 공식, 기호로 정리하면 <표 6-9>와 같으며, 이들 기호에 패싯 기호를 결합하면 J,381;421:5.42'N7의 분류 기호가 합성된다.

<표 6-9> CC의 분류 용례

	범주	패싯공식	분류기호
	기본범주	BF(기본패싯)	J
1970년대	시대범주	T	N70
일본	공간범주	S	42
벼	개체범주	1P1	381
바이러스	재료-성질범주	1MP	423
박멸	에너지범주	1E	5

6.4 전자 정보원 분류와 자료 분류법

6.4.1 개요

분류표의 개발에는 전문 지식과 추론 능력, 주제에 대한 이해와 더불어 많은 시간과 노력이 수반되며 이런 점에서 인공 지능이나 전문가 시스템으로 대치하는 것이 쉽지 않다. 비록 대량으로 출현하는 인터넷 자원에 대한 지식 구조화의 문제와 관련되어 전통적인 분류표를 이용하는 것이 부적합하다는 지적도 있으나 이들의 범주 구조와 계층 구조가 분류 및 검색에 유용한 것으로 밝혀진 바 있다.

웹의 초기에는 국제 십진 분류법(UDC: Universal Decimal Classification) 및 DDC와 같은 분류표가 많은 주제 게이트웨이에서 사용되었는데 전반적으로 이러한 역할이 쇠퇴하게 되었다. 이는 예산 부족, 웹의 빠른 성장 및 인터넷 자원의 역동성, 색인 형식 변화, 주제 게이트웨이와 포털의 협력 및 병합 등의 이유가 있었다.

웹자원의 조직 도구로 기존의 분류표를 적용하여 수행되었던 프로젝트를 포함하여 현재 운용되고 있는 사이트의 예는 <표 6-10>과 같다.

<표 6-10> 분류표를 적용한 웹 사이트의 예

	해당 주제 게이트 웨어(사이트)
DDC	• ADAM: Art, Design, Architecture & Media Information Gateway • Blue Web'm Browse by Subject Area • BUBL LINK(Univ. of Strathclyde) • Canadian Information By Subject(NLC) • Cooperative Online Resource Catalog(CORC) • CyberDewey • Internet Public Library Online Texts Collection • The Internet Resource(Napier University) • New Athenaeum: Internet Resource Guide Developed by Libraries • PICK: Quality Internet Resources in LIS • Full Text Documents(Thomas Parry Library) • Science Net: Subject(Toronto Public Library)
UDC	• Directory of Networked Resources: UDC "Shelfmark" Order • GERHARD(German Harvest Automated Retrieval and Directory) • WWW Subject Tree of WAIS Databases • SOSIG(The Social Science Information Gateway)
LCC	• Cooperative Online Resource Catalog(CORC) • CyberStacks(sm) • ICRC(Internet Collegiate Reference Collection) • Ready Reference Using the Internet • Web Resources Arranged by the LCC System
Facet Classification	• Epicurious • FLAMENOC(FLexible Access to MEtadata in NOvel Combinations) • LawforWA

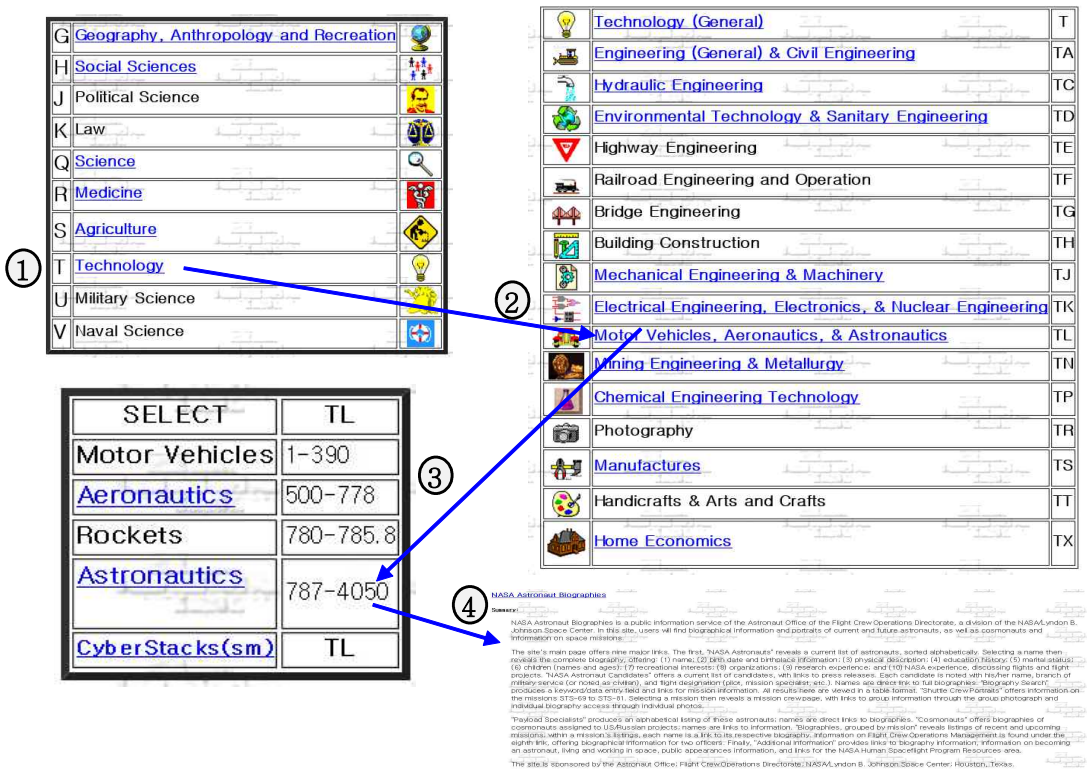
이들 대부분의 서비스는 키워드 탐색과 같은 다른 기능들과 함께 주제에 의한 브라우징 검색을 위해 분류표가 사용되었다. 이러한 서비스들은 상업 탐색 엔진에 비해 범위가 한정적이며, 실제 탐색 엔진이라기보다는 계층 구조와 삼호 참조 구조를 사용하는 탐색 디렉토리나 목록으로 간주하는 것이 더 적합할 수 있다. 다음은 전자자원에 대한 자료 분류표 적용의 사례를 살펴본 것이다.

6.4.2 CyberStacks[2)]

CyberStacks(sm)은 1998년에 아이오와 주립 대학에서 시작된 프로젝트로서 주요 WWW 및 기타 인터넷 자원을 대상으로 LCC를 이용해 주제를 범주화시켜 컬렉션을 구성하였다. 현재 해당 사이트는 프로토타입 시범 서비스이며 자연 과학 및 기술 과학 분야에서 선정된 중요한 인터넷 자원을 범위로 한정하고 있다.

각 자원에 대한 검색 방법은 계층 구조에 따라 검색 범위를 축소해 가는 브라우징 검색과 주제어와 분류기호에 대한 색인(Index)을 제공하여 접근하는 방법으로 구분된다. <그림 6-2>는 인터넷 자원을 일차적으로 포괄적인 상위 항목 수준으로 분류하고 점차 그 하위 수준으로 전개하고 있는 것을 보여주고 있다.

2) http://www.public.iastate.edu/~CYBERSTACKS/



<그림 6-2> LCC를 적용한 CyberStacks의 브라우징 과정

예를 들어, 'T 기술'을 주류에서 선택하면 그 하위는 'TL 자동차, 항공학, 우주항행학'으로 세분되고, 마지막으로 구체적인 분류 범위에서 'TL 787-4050 우주 항행학'이 제공된다. 이 항목에 분류된 자원에는 '미국항공우주국 우주비행사 전기'가 있다. 각 자원마다 간략한 요약이 제공되며, 필요한 경우에는 해당 자원을 사용하는 것에 대한 지침도 있다.

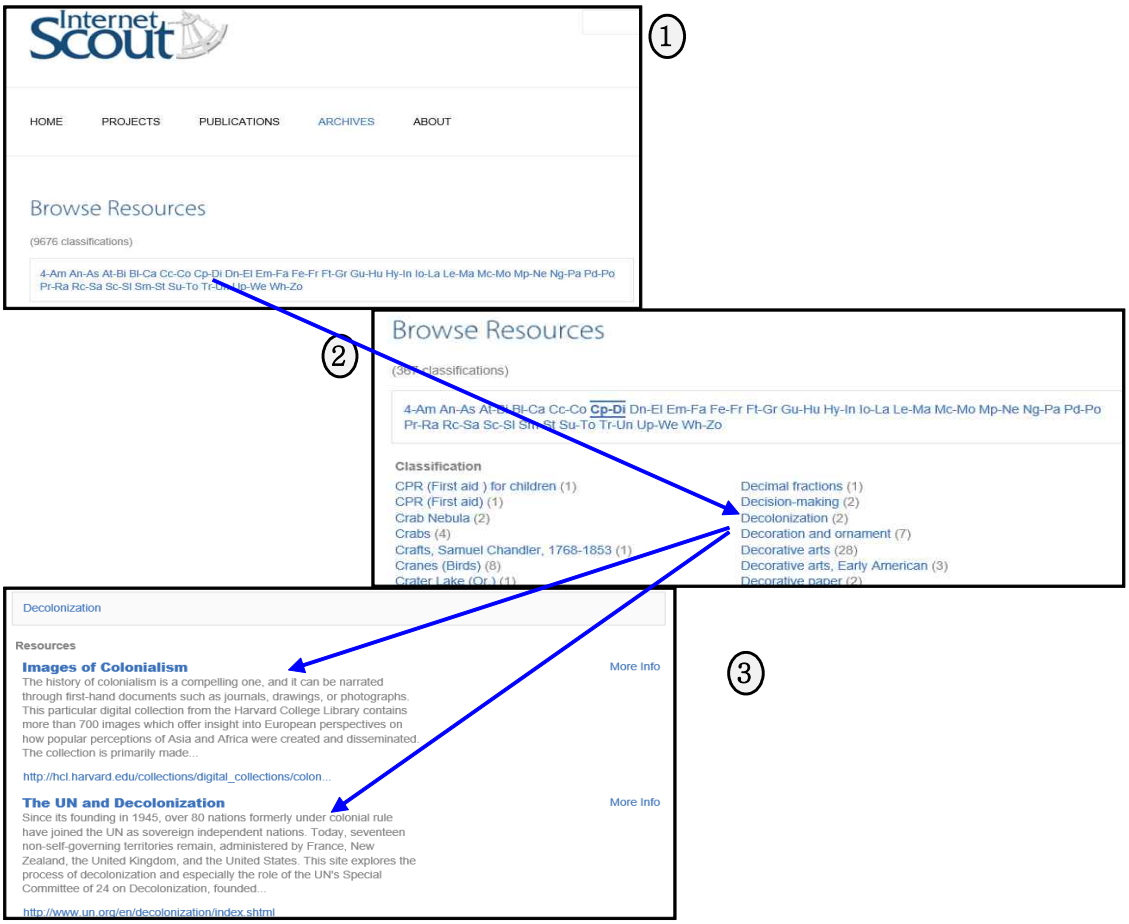
6.4.3 InternetScout Archives[3)]

InternetScout Archives는 위스콘신 대학의 컴퓨터 공학과에 속해 있으며 미국의 회도서관의 주제명 표목표 LCSH을 적용하여 약 9년간의 스카우트 리포트를 데이터베이스로 구축하였다. 현재 InternetScout Archives에는 엄선된 사이트에 대한 2만 5000개 이상의 비평적 주석을 포함하고 있다.

자원의 키워드 탐색은 표제, 저자, 발행처, LCSH와 같은 필드에 대해 상세 검색

3) https://scout.wisc.edu/archives/

이 가능하다. 자모순 LCSH 브라우징 검색의 경우 메뉴에서 검색하고자 하는 주제어의 알파벳 두문자를 선택하면, 해당 문자로 시작하는 주제명 목록이 제시된다. 예를 들어, 'Cp-Di' → '탈식민지화(decolonization)'를 순서대로 선택하면 해당 주제어 아래에 2건의 자원이 설명과 함께 접근가능한 사이트 주소가 제시된다(<그림 6-3> 참조).



<그림 6-3> LCSH를 적용한 InternetScout의 브라우징 과정

7. 시소러스

7. 시소러스

7.1 개요

문헌의 주제를 색인어로 변환하는 방법에는 기본적으로 디스크립터(용어)를 사용하는 시스템과 주제 간의 관계를 체계적으로 배열한 분류표의 두 가지가 있다. 그 중 디스크립터 시스템은 그 성질에 따라 전조합시스템(주제명표)과 후조합시스템(시소러스)이 있다. 출현배경에서 보면 주제명표는 그 기원을 카드 목록과 인쇄 목록에 둔 것임에 반해 시소러스는 자동화 시스템에 기반한 것이다(김태수 2000, 278).

시소러스에 대한 필요성은 방대한 정보를 운영하는 도서관, 언론사, 정부기관 등을 중심으로 대두되기 시작했다. 과거의 수작업으로 색인하는 방법을 탈피하여 개발된 자동색인 소프트웨어에서 동일한 개념이 다양한 용어로 표기되어 재현율을 낮게 만드는 문제점이 나타남으로써 이를 해결할 방법을 모색하게 된 것이다.

최근의 정보기술 혁신은 시소러스의 기능에도 많은 변화를 가져왔다. 초기의 시소러스는 수작업 서지 데이터베이스에서 사용되었다면, 현재는 전문(full-text) 데이터베이스가 주축을 이루는 환경으로 변화되었다. 또한 정보 이용자들의 이용 환경도 Windows를 기본으로 하는 GUI(graphic user interface) 환경으로 바뀌고 있다. 그러나 정보량의 증가에 따라 검색에서 발생하는 노이즈가 증가하게 되어 보다 적합한 정보를 원하는 이용자에게 많은 어려움이 되고 있다. 보다 효율적인 정보 검색을 위해서는 반드시 시소러스가 가미된 자연어 정보 처리가 요구된다. 따라서 시소러스에 대한 요구는 늘어날 것이며, 이에 따라 시소러스를 구축하여 보다 양질의 정보를 제공하려는 정보 서비스 기관이 증가할 것으로 보인다. 또한 정보검색 분야에서 정보 구축/관리(DBMS), 전문가 시스템 개발 분야, 인공지능 개발 분야 등 대량의 정보를 운영하거나 정보의 추론, 유추가 필요한 신기술과 상품 개발 분야에서 또한 사용자 위주의 환경에 필요한 자연언어 처리 분야 등에서 시소러스에 대한 수요가 증가할 것이다(최석두 2000, 5).

지난 수십 년간 정보센터 및 도서관에서도 분류표에서 시소러스로 전환하여 왔는데 이와 같은 변화의 주요 두 가지 이유는 첫째, '단어 시스템(word-systems)'이 사용하기 더 쉽기 때문이다. 즉 사용을 위해 어려운 리스트에 있는 기호를 찾아볼 필요가 없으며 단지 주제명을 선택하여 시도하면 된다. 둘째, 시소러스는 후조합 검색을 가능하게 하므로 이용자의 관점에서 자연어에 의한 주제어 검색이 용이하기

때문이다(UDC 1990, 88).

7.2 시소러스의 정의와 기능

7.2.1 시소러스의 정의

시소러스는 사전이나 백과사전과 같이 지식의 창고나 보고를 의미하는 말인데 일반적으로 '후조합 검색을 위해 설계된 자연 언어의 통제 어휘집', 혹은 '상위 및 하위 개념간의 관계를 제시하기 위하여 공식적으로 조직되고 제어된 색인 어휘집(ISO 1987)' 그리고 '구조적인 측면에서 보면 의미적, 종속적으로 관련된 용어를 제어해 둔 동적인 어휘집' 등으로 정의되고 있다. 이들 정의에서 볼 수 있듯이 시소러스는 일반적으로 '특정 학문 영역에서 사용되는 전문 용어 간의 관계를 의미 구조에 따라 조직한 전문 용어집으로서 어휘를 의미유형별로 분류하고 각 유형마다 동류의 용어를 배치함으로써 기술과 참조에 용이하도록 편찬된 어휘집'이다(김태수 2000, 277).

한편 김경호는 자연어 시스템의 경우 같은 주제라도 문헌 생산자나 색인 작성자, 이용자 간에는 그 표현하는 용어가 달라질 수 있어 문헌의 분석이나 색인 작성 과정에 많은 어려움이 야기됨에 따라 필요한 정보를 찾으려고 하는 이용자는 하나의 검색어만으로 해당 주제를 모두 검색할 수 없으므로 그 검색어에 관련된 개념의 대 · 소, 관련어 등을 모두 검색하여야 하는 불편이 뒤따르게 된다고 하였다. 이에 '그 해당 주제 분야에서 필요한 모든 개념을 수집하여 이들에 대한 개념의 대 · 소관계나 동의어, 동형이의어, 관련어 등을 적절히 조절하여 정보 시스템과 문헌 생산자, 색인 작성자, 이용자 간에 통일적으로 사용할 수 있도록 통제하여 둔 용어 제어표'를 시소러스라고 하였다(2002, 264-265).

시소러스는 문헌의 주제를 디스크립터(용어)를 사용하여 색인어로 변환하는 시스템이며 처음부터 분류표의 사용을 전제하지 않은 자동화 시스템에 기반한 것이다. 그러나 대부분의 시소러스에서 검색의 보완용으로 분류 구조를 수용하고 있는데 이것은 바로 개념이 체계적으로 배열되지 않음으로 해서 오는 한계를 보완하기 위한 것으로 이용자에게 별도의 검색 장치를 제시하기 위한 것이다(김태수 2000, 278).

7.2.2 시소러스의 기능과 목적

시소러스의 기본적인 목적은 특정한 용어의 선정은 물론 그와 관련된 용어로 질

의를 확장하여 재현이라는 측면에서 검색의 효율을 개선하는데 있다. 구체적으로 시소러스는 다음과 같은 기능 및 목적을 갖는다:

• 특정 주제 분야의 지식 구조를 보여준다.

• 색인자와 이용자가 사용한 자연 언어를 각각 색인 작성과 검색에 사용할 통제 어휘로 번역하는 수단을 제공한다.

• 색인자를 돕는 측면으로 문헌의 주제를 나타내는 색인어 부여시 일관성을 유지하게 한다. 즉, 문헌에 대한 주제 분석을 통해 얻어진 주제 개념들을 그 시스템이 사용하고 있는 색인 표목으로 변환시키기 위한 표준 어휘를 제공한다.

• 용어 간의 참조 체계를 통해 의미 관계를 지시함으로써 용어의 확장과 축소를 가능하게 한다.

• 문헌의 탐색 시 탐색 보조 도구가 된다. 이용자의 관점에서 적절한 탐색어를 선택하고 축소하여 문헌의 탐색을 용이하게 하고 정확하게 한다.

• 새로운 개념을 기존 개념들 간의 관리 체계에 맞추어 제 자리를 잡게 한다.

7.3 자연어와 통제 어휘

자연 언어를 사용하는 온라인 정보 시스템이 출현하고 있으나 자연 언어는 통제 어휘에 비해 장점과 동시에 약점을 지니고 있다. 일반적으로 자연 언어가 검색에서 지니는 장점으로는 최신성과 특정성을 지니고 있어 검색의 정확성을 개선할 수 있으며, 인명이나 기관명과 같은 고유명의 검색에 효과적인 점을 들 수 있다. 그리고 자연언어의 망라성으로 인해서 재현율이 높으며 저자가 사용한 용어를 사용하므로 색인에 오해의 여지가 없다. 또한 입력 비용이 절감되며 데이터베이스간의 자료교환에 효과적이다(김태수 2000, 281).

한편 다수의 동의어와 하위 개념이 있을 때 탐색에 부담이 되고, 구문(용어간의 부정확한 관계)의 문제로 인해 부적합 문헌의 검색 가능성이 높으며, 용어의 망라성으로 인해 검색의 정확도에 부정적 영향을 줄 수 있다.

따라서 특정 개념에 대해 가능한 모든 용어를 기억하기 어렵고 특히 주제와 친숙하지 않은 경우에는 통제 어휘를 사용할 필요가 있다. ‘통제된(controlled)’다는 것은 언어가 문헌의 내용을 일관성 있게 기술하기 위해 제공되는 특수한 특징을 갖는 것을 의미하며 다음과 같은 여러 가지 방법으로 통제될 수 있다:

• 자연어에서 언어 변화를 제한하는 방법(예: 복수, 단수, 철자 등)

- 어구를 형성하는 단어의 순서와 구조를 규제하는 방법
- 동의어 및 유사 동의어의 수를 줄이는 방법
- 범위 주기에 의한 용어의 의미를 설명하는 방법
- 문맥을 제공하는 수단에 의한 용어의 의미를 설명하는 방법
- 계층적 관계 및 다른 관계와 같이 용어 간의 관계를 제시하는 방법

통제 어휘의 사용이 검색에 주는 긍정적인 면은 시소러스의 의미 구조를 통해 탐색 전략을 조정할 수 있으며, 동의어와 유사 동의어에 대한 조기성으로 인해서 재현율이 효과적이라는 것이다. 또한 특정 개념을 이와 유사한 용어로 연결하여 탐색의 부담이 적으며 복합어와 동형 이의어의 통제를 통하여 검색의 정확성을 높일 수 있다.

그러나 용어의 특정성이 낮고 망라성이 결여되어 있으며 입력 시 오류의 가능성이 있다. 아울러 용어를 수록하는데 많은 시간이 소요되고 저자가 의도한 용어를 잘못 사용하거나 인위적인 용어를 사용하므로 오해가 있을 수 있고, 자연 언어와 비교하여 입력 비용이 크고 호환성이 적어 데이터 교환에 장애가 될 수 있다.

시소러스를 구축할 때 고려해야 할 사항은 용어의 특정성을 높여서 검색의 정확률을 향상시키면 재현율의 저하를 가져오며 반대로 어휘의 특정성을 낮추면 재현율은 향상되나 정확률에 손상을 준다는 것이다. 아울러 통제 어휘나 자연 언어간의 상호보완적인 성격도 시소러스 구축 시 고려되어야 한다.

용어를 체계적으로 연결하고 특정성 있는 용어와 복합어를 사용하여 고도로 구조화된 복잡한 시소러스를 구축하는 것은 입력과 관리에 많은 비용을 초래하게 되고 또, 이러한 수준의 시소러스를 구축하기 위해서는 검색 결과의 개선을 예상할 수 있어야 한다. 반대로 포괄적인 용어와 최소한의 구조로 구축된 시소러스는 검색 결과가 불만족스럽고 이로 인해 오히려 더 많은 비용을 초래할 수 있다(김태수 2002, 282).

시소러스를 만드는 작업은 계속적으로 갱신해야 하는 반복적인 작업으로 노동 집약적이며, 시간과 노력, 그리고 비용이 소요되는 작업이다. 그러므로 이를 위해서는 표준화되고 잘 정의된 기술 규칙이 필요하다. 명확히 정의된 기술 규칙이 마련된다면 개별 기관이 수행하던 개별 규칙을 통합하여 보다 쉽게 통일된 데이터를 구축할 수 있게 되며, 용어 데이터의 생산을 활성화하게 되고 검색 효율의 향상을 가져오게 되어 유형, 무형의 경제적 이점을 가져올 수 있다(한유석, 설근수 2004, 196).

7.4 시소러스 개념 간의 의미 관계

7.4.1 동의 관계

동의 관계(USE, UF)는 복수의 용어가 동일한 의미를 지닐 때 성립되며, 동의어는 개념적 의미에서 의미 성분이 동일하므로 상호 함의관계(mutual entailment)를 가진다. 그런데 시소러스는 검색을 목적으로 하기 때문에 용어의 의미를 의도적으로 제어하고 있다. 동의 관계에 있는 용어 중 특정 용어를 디스크립터(우선어)로 선정하고 다른 용어는 비디스크립터(Use for: UF)로 취급되어 이차적인 접근점을 사용한다.

동의어의 유형으로는 표준어와 방언, 격식, 우리말과 외래어, 전문성, 내포, 학술명과 제품명, 상이한 철자, 경쟁 관계의 용어, 통용어와 고어, 약어와 완전명 등이 있으며, 유사 동의어를 포함한다.

7.4.2 계층 관계

개념의 의미 구조를 계층 관계로 표현한 것으로 한 쪽이 의미상 다른 쪽을 포함하거나 다른 쪽에 포함되는 관계이다. 의미의 포함 관계에는 1) 개체와 그 개체가 속한 부류 포함관계(class inclusion), 2) 신체 구조나 지역, 인공물(구조물, 기계 등)과 같이 자연계에 속하는 개체 및 대학이나 기업체, 정부 조직과 같은 구조적-공간적 관계에 기초한 부분-전체 포함관계(part-whole inclusion)와 지식 영역간의 주제 포함관계, 3) 사례 관계가 있다.

일반적으로 계층 관계는 BT, NT로 표현되며 때로는 이 관계를 더 전개하여 속종관계는 BTG/NTG로, 부분-전체 관계는 BTP/NTP로 사례 관계는 BTI/NTI로 나타내기도 한다.

7.4.3 결합 관계

결합 관계에는 다음과 같은 유형의 관계를 포함한다;

• 대등 합성어: 구성요소 A, B가 대등한 자격으로 결합되어 하나의 개념을 이루는 것을 말한다. 그러나 시소러스는 특정 학문 영역을 대상으로 한 전문 용어를 대상으로 한다는 점에서 대부분의 시소러스에서 관용적 의미를 지닌 대등 합성어에

대한 구조화에 대해서는 그다지 관심을 기울이지 않는다.

• 연관 관계: 분석된 개념의 의미를 제시하고 그 개념과 관련된 다양한 관점을 지시하기 때문에 검색에서 중요시된다. 일반적으로 연관 관계는 계층 관계나 대등 관계 이외의 용어 관계를 의미하며 검색이라는 의미에서 주로 정의된다. 시소러스 이용자가 공유하는 지시의 틀에 따라 관계를 설정할 수 있으며 주로 이용자의 질의나 그 주제 영역의 문헌을 분석하여 연관 관계를 확인할 수 있다.

• 반의어: 경험과 판단을 대립 관계로 파악하려는 인간의 보편적인 성향에서 비롯된 것으로, 일반적으로 연속체상의 개념 영역을 상호 대립적이고 배타적인 두 영역으로 양분하는 용어로서 검색 시 강하게 연상되는 용어를 말한다.

7.4.4 복합 관계

하나의 개념이 둘 이상의 복합적 의미 관계를 가질 때 다의 관계가 성립되며 다의어는 기본 의미와 파생 의미로 구성된다. 다른 복합 관계는 동형이의어로 하나의 표현에 여러 의미가 대응되는 복합적 의미 관계로서 다의어와 명확한 구분이 어렵다. 일반적으로 시소러스에서는 동형이의어로 인해 야기되는 모호성을 해소하기 위해 다의어를 복수의 디스크립터로 제시하고 그 의미를 한정하기 위한 한정어를 도입하고 있다.

7.5 분류와 시소러스

분류는 자료의 물리적 배열은 물론 시소러스에서 디스크립터의 의미 관계를 결정짓는 중요 수단으로 그 유용성이 인정되고 있다. 시소러스에서 디스크립터를 구조화할 때 대개 대상 용어를 크게 범주로 구분한 다음 각 범주 내에서 개념 관계를 설정해나간다. 이때 범주의 구분 기준은 기존의 주제 분류나 학문 분류표 또는 패싯을 사용한다. 일반적인 방법은 학문 분야나 주제 분야로 구분하고 그 안에서 패싯에 따라 구분하는 경우가 많으나 대구분부터 패싯을 적용하는 경우도 있다. 따라서 분류와 시소러스는 지식의 구조화를 위한 동일한 과정을 거치며 표현 형식에서 차이가 있다.(김태수 2000, 278).

최근 이용자의 검색 환경을 고려하여 개발된 탐색용 시소러스는 이용자의 정보요구를 유연하고 다양하게 수용할 수 있어 주제 분석에 기여할 수 있으며, 이런 점에서 시소러스는 지식의 구조화 과정에서 분명 하나의 진전으로 평가할 수 있다. 기존의 전통적인 분류 체계 및 주제명 표목표와 비교할 때 시소러스는 다음과 같은

특징을 지닌다(한유석, 설근수 2004, 189-195)

• 다양한 분류 체계의 수용: 일반적으로 분류 방법은 고착적이기 때문에 상황에 따라 분류 체계를 융통성 있게 표현할 방법이 없으나 불특정 다수인 보통의 정보이용자는 다양성 속에서 나름대로의 유사성을 파악하고 있으며 자신의 분류체계를 통하여 사물을 보고 있다. 일반적으로 특정의 분류 체계가 다른 분류 체계를 수용하는 것은 매우 어려운 반면에, 시소러스는 다양한 분류 체계를 완전하지는 않지만 상당부분을 수용할 수 있는 장점을 가지고 있다.

• 복수의 상위 개념: 하나의 개념은 여러 가지 측면에서 볼 수 있는데 어떤 개념이 여러 가지 그룹(집합)에 속하고, 각 그룹을 대표하는 개념이 있게 되면 복수의 상위 개념을 갖게 된다. 이는 매우 자연스러운 일이며 많은 용어가 복수의 분류 기호를 갖게 된다. 문헌 분류 체계에서도 두 가지 이상의 분류 기호를 가질 수는 있으나 실제의 분류에서는 그렇게 흔한 일은 아니다. 그러나 시소러스에서는 두 개의 상위 개념어와 두 개의 하위 개념어를 갖는 구조의 용어 관계를 허용함으로써 용어 관계 표현을 훨씬 풍부하게 하며 관점의 유연성을 배가시켜줄 수 있다.

• 세목의 배제: 세목을 배제함으로 복잡한 조합 및 분류 규칙의 애매함을 배제하고 조합의 융통성을 보장할 수 있다. 즉 시소러스에서는 용어 순서에 대한 애매함이 없어지며 조합이 자유로워진다.

• 느슨한 범주화: 시소러스에서는 '속하는가, 속하지 않는가'를 표현하는 집합에서 '어느정도 속하는가'라는 모호한 집합도 표현할 수 있도록 관련 개념어가 느슨하게 연결되어 있다. 이용자는 이 관계 간의 경로를 적절히 이용함으로써 자신의 관점에서 필요한 정보를 얻을 수 있게 된다.

• 개념 구조와 특정성 조절: 범주 분류 체계에서 개념의 특정성을 강조하기 위해서는 계속적으로 세목을 만들어 가야 한다. 그러나 모든 세목의 전개는 일정한 틀을 가져야 한다. 그렇지 않으면 전개된 세목이 한 번의 특정한 경우에만 사용되고 마는 일이 생기게 된다. 따라서 범주 분류 체계는 만들기도 어려우며 구조의 변경 또한 어렵다. 그러나 시소러스형 용어에서는 개념의 특정성을 복합명사나 명사구를 이용하여 다양하고 자유스럽게 표현할 수 있으며 이용자가 갖고 있는 개념의 특정성에 연동할 수 있다. 뿐만 아니라 새로운 개념이 추가되면 개념 구조는 바뀔 수도 있다. 특히 색인의 문제와 검색의 문제는 색인 작업 측면에서는 밀접한 관련을 갖고 있으나 용어 관리 상으로 보면 서로 상당히 독립적이다. 개념 구조가 바뀌었다고 색인어를 꼭 바꾸어야 하는 것은 아니다.

• 용도의 일반화: 시소러스는 색인과 검색에 사용된다. 그러나 색인 기능과 검색

기능의 수준은 일정한 것이 아니라 데이터의 종류, 양, 분야, 이용자의 요구 등에 따라 천차만별이다. 전분야의 전문 데이터에 대한 자동 색인과 검색을 대상으로 하는 시스템에서는 수백만 단위의 어휘를 필요로 할 수도 있다. 결국 용어는 어느 용도에나 적용할 수가 있어야 할 것이며, 새로운 요구에도 적응할 수 있어야 할 것이다.

7.6 시소러스의 발전

종래의 고전적인 범주화에 따른 개념 체계의 설정보다는 이용자의 인지 행동과 개념 인식 과정이 중요시되고 이것은 지식의 구조화에서 광범위하면서도 고도의 의미 처리 기능을 지닌 자연 언어 처리 기술이 기초가 되어야 한다는 것을 의미한다. 실제로 시소러스는 전문가 시스템이나 인터페이스 시스템, 객체지향 시스템, 하이퍼텍스트 시스템, 기계 번역, 자동 초록 등의 여러 분야에서 그 응용 영역이 확대되고 있다(Schmitz-Esser 1991, 143). 이에 따라 특정 영역에 국한되지 않고 다양한 분야의 데이터베이스에 사용될 수 있는 시소러스의 개발이 이루어지고 있다.

시소러스의 개발을 위한 향후 과제로는 시소러스의 수록범위 확장과 상호참조 이외에 디스크립터의 의미 구조를 다양하게 표현하기 위한 시도가 필요하다. 일반 어휘를 대상으로 한 어휘집의 개발이 진행됨에 따라 시소러스와 일반 어휘집과의 경계가 점차 불분명해지고 있다(김태수 2000, 351).

한편 시스템에서 사용하는 지식 베이스에서도 상호 작용하는 인공 지능 에이전트의 등장으로 인해 지식 구조의 호환성에 따른 문제가 제기되고 있고 이에 따라 온톨로지에 관한 연구가 활발히 이루어지게 되었다. 온톨로지는 용어의 의미 관계와 연결 정보를 보다 유동적이고 상세하게 기술하기 위한 시소러스의 확장 개념으로 볼 수도 있다. 예를 들어 용어 온톨로지는 시소러스와 의미망의 연계를 통한 영역이라고 할 수 있다. 정보 처리 기술과 인공 지능 기술을 응용하고 아울러 개념의 이용 장면을 중시하는 실제적인 이유에서 이러한 변화는 당연한 귀결이라고 평가된다.

8. 시소러스 구축 사례

8.1 개요

8.2 유네스코 시소러스

8.3 미국 의회도서관 주제명 표목표

8.4 시소러스와 결합된 분류표

8. 시소러스 구축 사례

8.1 개요

Thesaurus Guide: Analytical Directory of Selected Vocabularies for Information Retrieval』에서는 시소러스의 주제 분야를 총류 일반(도서관, 과학기술), 정보 영역(정보학, 커뮤니케이션, 컴퓨터), 수학과 물리화학(수학, 물리학, 엔지니어링, 기계 공학, 전자 기술, 에너지, 원자력), 물리화학 기술(화학, 야금술), 천문학 및 기하학(천문학, 기하학, 지리학, 광물학), 농학과 영양학(농학, 식물학, 임학, 영양학), 생물의학(생물학, 의학, 수의학), 종교와 환경 과학(환경, 도시 공학, 종교, 교통), 사회 과학(사회학, 심리학, 교육학, 경영학, 정치학, 법학, 경제학, 군사학), 문화와 예술(문화, 역사, 인종학, 박물관학, 미술, 언어학, 음악, 스포츠)의 10개 범주로 구분하고 있다(최석두, 2000).

<표 8-1> 분야별 시소러스 구축 사례

	시소러스명	구축기관
총류 일반	UNESCO Thesaurus	UNESCO
	LCSH	Library of Congress
정보영역	Root Thesaurus	British Standards Institution
과학기술	INSPEC Thesaurus	The Institution of Electrical Engineers
	JICST 科學技術用語ツソーラス	日本科學技術スンター
	TEST(Thesaurus of Engineering and Scientific Terms)	Engineers Joint Council
	NASA Thesaurus	NASA
	INIS Thesaurus	IAEA
생물의학	MESH(Medical Subject Heading)	National Library of Medicine
종교 및 환경과학	CIT(Construction Industry Thesaurus)	CIT Agency
	EI Thesaurus	Engineering Information inc.
사회과학	Thesaurus of ERIC Descriptor	James E. Houston
	TSIT(Thesaurus of Sociological Indexing Terms)	Barbara Booth and Michael Blaired
문화 예술	A Women's Thesaurus	National Council for Research on Women

지금까지 개발된 시소러스 중에서 개별 주제 분야에서 용어수가 많고, 인지도가 있는 것을 중심으로 주제별로 구분하면 <표 8-1>과 같다. 그 중 몇 가지 사례를 중심으로 다음과 같이 구체적으로 살펴보았으며, 패싯 분류표를 적용하여 시소러스와 결합한 사례를 포함하였다.

8.2 유네스코 시소러스(UNESCO Thesaurus)[4)]

유네스코 시소러스는 유네스코 도서관에서 사용되는 정보의 색인과 탐색을 위해 계층적으로 구조화한 통제 어휘 리스트이며, 도서관뿐만 아니라 영국의 기록관(UK Records Office)이나 PMB(웹기반의 오픈 소스 통합 도서관 시스템)과 같은 여러 시스템에 적용되고 있다. 1977년에 초판이 발행되고 1995년에 업데이트된 개정판이 발행되었으며 2000년에 영어판의 온라인 버전이 나오게 되었다. 2003년 이후 프랑스어, 스페인어, 러시아어 온라인 버전이 매년 추가되었다.

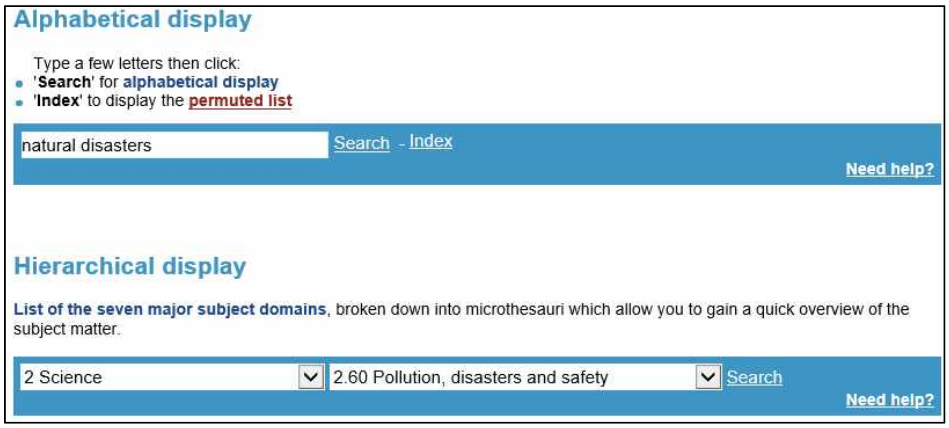
이 시소러스의 목적은 유네스코의 콘텐츠를 기술하는데 주제 분석의 일관성을 부여하고 온라인 목록으로부터 적절한 정보의 탐색을 용이하게 하려는 것이다.

이 시소러스의 특징은 대상 범위가 유네스코와 관련하는 학제적 지식 분야를 모두 포괄하고 있다는 것이다. 유네스코에서 발행하고 수집하는 문서 및 출판물의 주제 분야는 교육, 문화, 자연과학, 사회 과학 및 인류학, 커뮤니케이션 및 정보, 국가명 등에 걸쳐 있다. 구조적 측면에서는 지리 정보 디스크립터를 제외하고 단일 계층 구조를 가지면, 영어, 프랑스어, 스페인어, 러시아어의 다국어 버전으로 이용가능하다. 또한 시소러스 구축의 국제 표준을 적용하여 ISO 2788과 ISO 5964에 따라 구조화하였다.

시소러스의 구조는 용어(terms), 용어 관계(term relations), 범위 주기(scope notes) 등으로 이루어져 있다. 시소러스의 용어 관계에서 일반적으로 사용되는 기호(SN, USE/UF, BT/NT, RT)이외에 MT(Microthesaurus)를 사용하는 것이 특징인데 이는 디스크립터가 속해 있는 하위 시소러스의 이름과 번호를 지시하는 것이다. 즉, 시소러스의 기본 범주는 7개의 주제 도메인(1.Education, 2.Science, 3.Culture, 4.Social and human sciences, 5.Information and communication, 6.Politics, law and economics, 7.Countries and country groupings)으로 구성되어 있으며 각각 범주 아래에는 88개의 하위 범주들로 세분되어 있다. 이와 같은 범주의 계층은 시소러스의 알파벳 키워드 검색과 더불어 브라우징 검색에 적용되고 있다(<그림 8-1>참조). <그림 8-2>는 'natural disasters'라는 용어로 검색을 수행

4) http://databases.unesco.org/thesaurus/

한 결과이며 용어에 관련된 다국어 표현과 MT, UF, BT, NT 등의 다양한 관련 용어들을 확인할 수 있다. 또한 각각의 용어 뒤에 표시되는 숫자는 유네스코 도서관의 온라인 목록으로부터 주제와 관련된 유네스코 문헌수를 알려주는 것으로 해당 서지리스트로 연결해준다.



<그림 8-1> 유네스코 시소러스의 검색 화면

<그림 8-2> 유네스코 시소러스 용어 검색 사례

현재 유네스코 시소러스는 4,500여개의 우선어가 구축되어 있으며 영어와 러시아어 7,000개 이상의 용어, 프랑스어와 스페인어 8,600개 이상의 용어를 포함하고 있다. 또한 640개 이상의 범위주기와 17,000개 이상의 용어관계를 포함하고 있다. 유네스코에서는 새로운 용어와 용어 사용의 변화를 수용하고 발전시키기 위해 지속적

으로 관리하고 있으며 2011년 이후 아랍어 번역, 새로운 시소러스 관리시스템 도입, 시멘틱 웹에서의 활용을 용이하게 하기 위한 SKOS로의 변환, 다른 시소러스와의 상호운용성의 모색 등과 관련된 프로젝트가 진행되었다.

8.3 미국 의회도서관 주제명 표목표(LCSH)

미국 의회도서관 주제명 표목표(Library of Congress Subject Headings: LCSH)는 1914년 초판이 발행되었으며 본래 미국 의회도서관 장서의 주제 접근을 위한 도구로 고안되었으나 지금은 미국 내 도서관은 물론 세계 여러 국가의 도서관에서 주제 검색 도구로 사용되고 있는 일반 통제 어휘집의 하나가 되었다.

LCSH는 지난 수십 년 동안 인터넷, WWW, 메타데이터 등 정보 검색 환경의 변화에 따라 1988년 제 11판부터 참조 구조를 비롯하여 외형적으로 큰 변화가 있었고 시소러스 방식의 표시 기호를 도입하였다. 또한 주제명의 수도 계속 늘어나 2013년에 인쇄본이 35판까지 발행되었으며 332,500건의 주제명을 수록하고 있다.

표목의 구조는 다음과 같이 주표목과 세목, 참조, 기타로 구성된다(정연경, 2013).

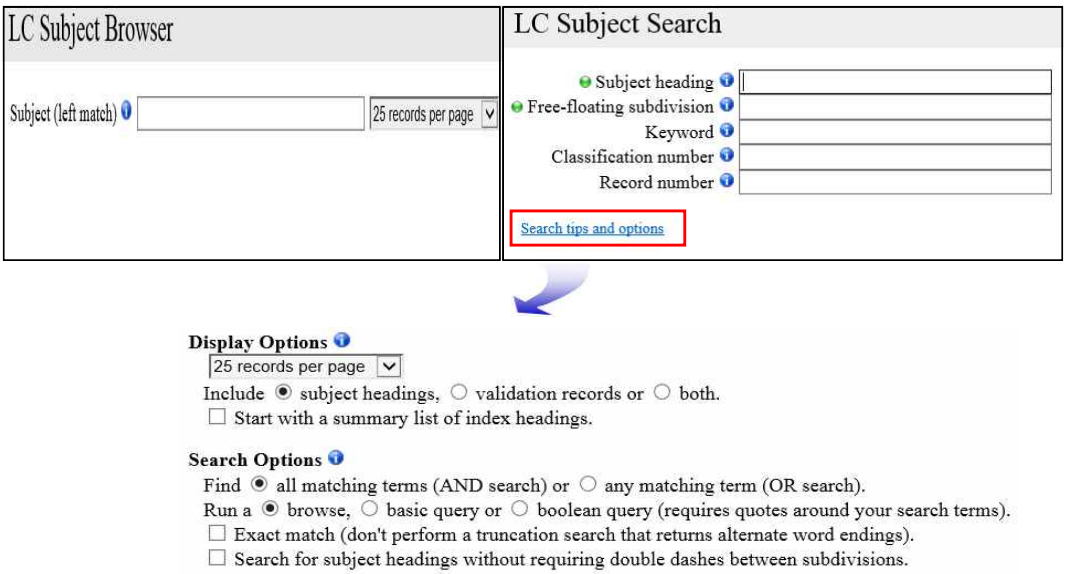
첫째, 주표목은 하나 이상의 단어로 이루어지고 두 단어 이상의 표목은 형용사 + 명사, 접속사(and) 연결 혹은 전치사구로, 도치 표목의 경우는 언어나 국적 등을 기술하는 형용사로 시작하는 단어이다. 복합어가 표목으로 지정되기도 하는데 2개 표목이 매우 유사하여 저작에서 함께 자주 다루어지는 경우이다. 한정어는 동음이의어나 용어의 문맥상에 고유한 의미를 명확히 설정하기 위해 원괄호 안에 기입한다. 지역 세구분 지시는 주표목에 지역 세목의 부가 여부를 지시하며, 원괄호 안에 ‘(May Subd Geog)’ 또는 ‘(Not Subd Geog)’로 표기한다.

둘째, 세목은 주제의 특수한 측면들을 표현하는 것으로, 표목의 구체성을 높이는데 사용되었으며 긴 줄표(-)로 식별한다. 성격에 따라 주제 세목, 형식 세목, 시대 세목, 지리 세목이 있으며, 특이하게 자유 부가 세목(free-floating subdivision)은 분류표의 조기표와 유사한 기능을 하는 것으로, 일부 주제와 형식 세목을 표준 집합으로 구성하여 가능한 표목에 모두 조합할 수 있다.

셋째, 참조는 동의어 관계(USE/UF), 계층관계(BT/NT), 연관관계(RT) 등 시소러스의 기본 관계를 일부 적용하고 있다. 용어 간의 관계성을 식별하는 참조 이외에도 분류기호나 범위 주기와 같은 용어의 부가 정보에 관한 참조도 있다. 분류 기호에는 LCC 기호를 부여하며, 각괄호([ ])로 식별한다. 현재 87,500건의 주제명에

분류 기호가 부여되어 있으며 범위주기는 목록 레코드에 주제명을 일관성 있게 부여하기 위하여 주제의 범주를 명시하고, 관련 표목들 간의 차이점을 기술하며, 표목의 몇 가지 의미 중에서 도서관 목록에서 사용이 제한되는 경우를 설명하는데 사용된다.

LCSH를 이용할 수 있는 웹 사이트로 Classification Web이 있으며 회원(구독) 기관만 접근 가능하다.[5] 해당 표목의 전거 MARC 레코드로 연결이 가능하나, 서지 데이터로 연결되지는 않는다. 주제명의 검색뿐만 아니라 브라우징 할 수 있는 기능이 추가적으로 구현되어 인터페이스가 상대적으로 이용자 친화적이다. 또한 기본 검색창 아래에 "Search tips and options" 링크를 통해 다양한 디스플레이 및 검색의 옵션을 선택하여 이용자가 원하는 방식으로 검색 및 출력을 할 수 있는 기능을 제공한다(<그림 8-3> 참조).



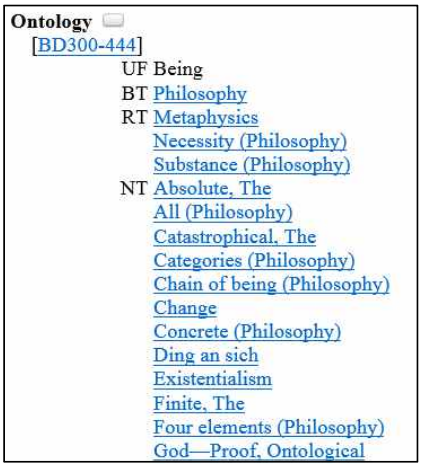
<그림 8-3> LCSH 브라우징/검색 화면과 검색옵션

실제 'ontology'라는 주제명 표목 검색을 실시한 결과 <그림 8-4>와 같이 LC 분류 기호, 동의어, 상위 관계어, 하위 관계어, 연관 관계어를 확인할 수 있다.

LCSH의 장점은 통제 어휘집인 LCSH를 활용하여 주제의 범주를 확장하거나 축소하여 탐색함으로서 어떤 한 주제명 표목에 관련된 모든 자료들을 신속 · 정확하게 찾을 수 있다는 것이다. 또한 더 넓은 주제어나 더 좁은 주제어의 관계를 탐색할 때 관련 주제어를 함께 브라우징할 수 있다는 점을 들 수 있다. 반면 단점으로는

5) http://classificationweb.net

이용자의 검색어와 LCSH의 표목에 사용된 색인어간의 차이로 검색의 효율성이 떨어지는 사례, 국가마다 역사적 시대 구분의 통일성이 없어서 역사적 측면을 반영하는데 문제가 있다는 것과 각 언어의 차이로 인해 표목명의 이해에 어려움이 있다는 것이다.



<그림 8-4> LCSH의 'ontology' 검색 결과 화면

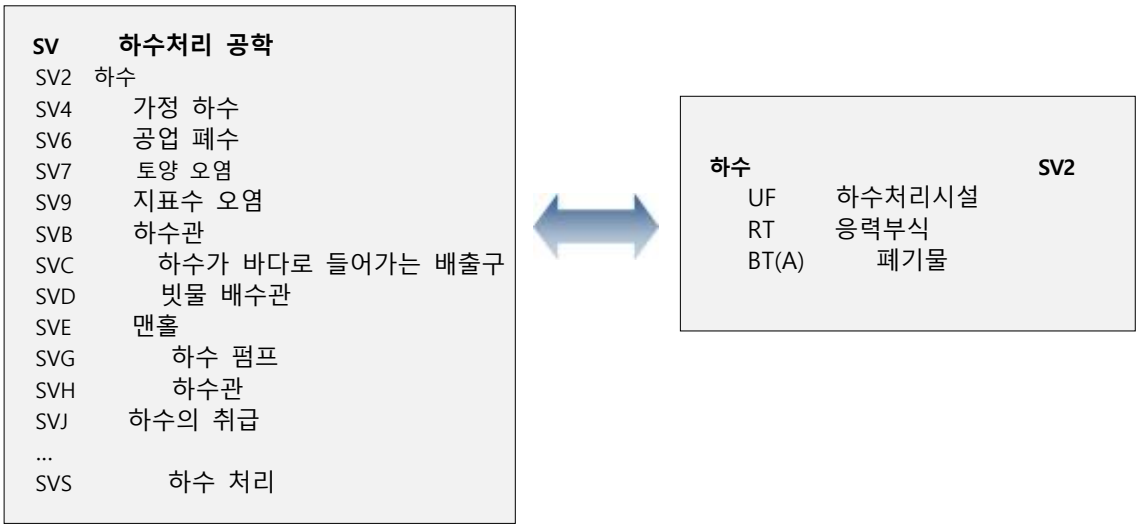
8.4 시소러스와 결합된 분류표

8.4.1 시소러패싯

시소러패싯은 공학 및 관련 분야의 패싯 분류표와 시소러스로서 영국의 전기 회사에서 사용하기 위해 개발된 것이다. 현재는 다소 시대에 뒤떨어지는 것으로 여겨지지만 분류표-시소러스 통합 시스템 유형의 시초가 된 분류표라는 점에서 의의가 있다. 특히 BSI Root Thesaurus는 시소러패싯을 기반으로 구축되었다.

시소러패싯에서는 분류표의 계층 구조에 제시된 관련어가 분류표의 색인에는 반복되지 않는다. 색인에서는 새로 추가된 관련어만 수록되기 때문에 분류나 자모순 주제색인을 위해서는 분류표와 색인을 함께 이용해야 한다. 예를 들어 <그림 8-5>는 '하수'에 해당하는 시소러패싯의 일부분과 색인/시소러스의 엔트리를 보여주는 것이다. 분류표에서 '하수'라는 용어의 상위어는 '하수처리 공학'이며 하위어는 '가정하수', '공업 폐수' 등이 있고, 연관어로 '하수관', '하수 펌프' 등이 있음을 확인할 수 있다. 이와 같이 분류표에 있는 용어들은 색인/시소러스에서는 반복되지 않고 있는 반면 본표에서 제시되지 않은 용어들을 포함하고 있다. 즉 색인/시소러스의 '하수'의 관련용어로 제시된 '응력 부식'은 분류표 상의 '하수 처리 공학'이 아닌

BT(A)에서의 '폐기물'이라는 추가된 상위어와 관련이 있음을 알 수 있다. 이와 같은 점은 분명히 경제적인 장치이기는 하지만 모든 용어를 확인하기 위해서는 본표와 시소러스를 모두 사용해야 한다는 것을 의미한다.



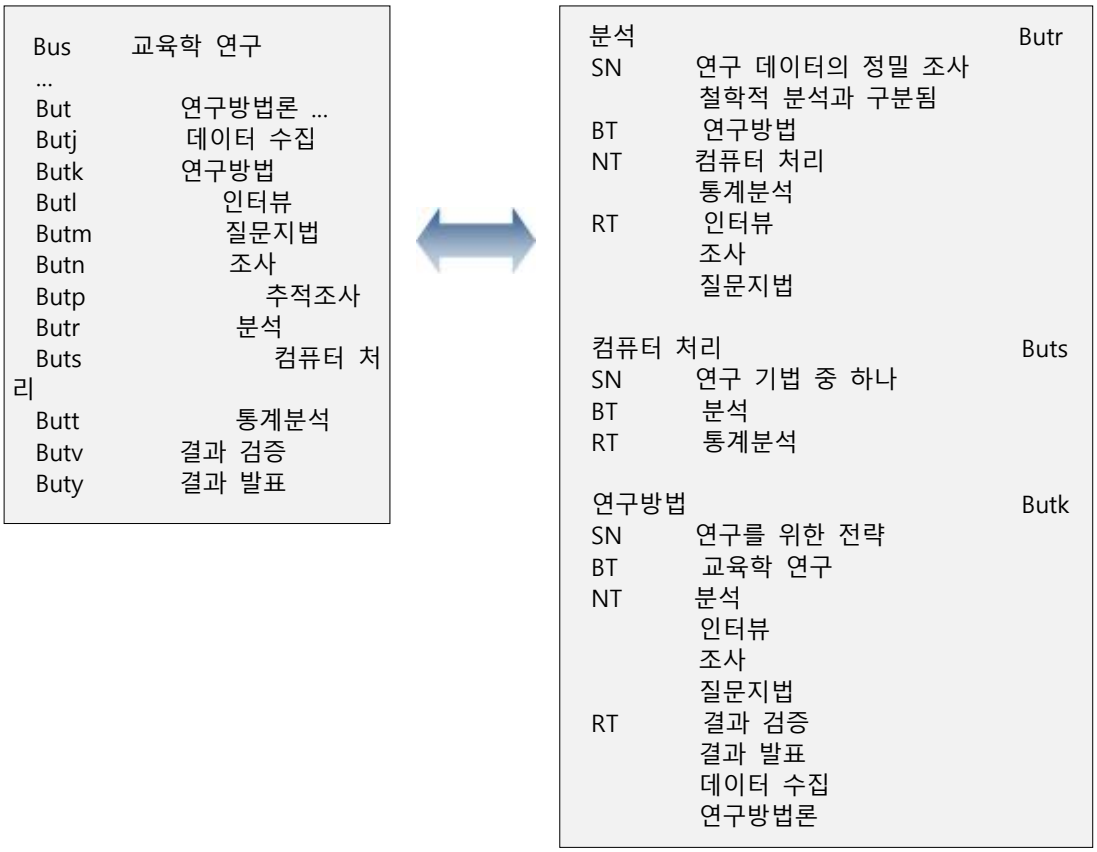
<그림 8-5> 시소러패싯 분류표와 색인/시소러스 사례

8.4.2 런던 교육학 분류법(LEC)

런던 교육학 분류법은 원래 런던 대학교의 교육 대학원 도서관에서 사용하기 위해 구축되었으며 현재도 사용하고 있다. 2판은 시소러패싯의 영향을 받은 반면, UNESCO 시소러스의 교육 부분에 영향을 미치고 있다.

런던 교육학 분류법은 시소러스 형식의 자모순 영역이 포함되어 있으며, 기호법은 알파벳 기호로만 이루어져 있어 순서를 나타낼 수 있고, 분류 기호의 대부분이 세 자리 기호로만 이루어져 있다. 따라서 이해하거나 쓰기 쉬울 뿐만 아니라 Hab(교육경영), Lob(시청각 도구)와 같이 기호들은 대부분 발음이 가능하다.

<그림 8-6>는 이 분류표에서 발췌한 색인/시소러스 엔트리와 이 엔트리를 기반으로 작성된 분류표의 사례를 보여준다. 즉 색인/시소러스 엔트리들은 분류표에서 직접적으로 획득되었으며, 그 분류표의 자모순 색인이라는 부가적인 역할도 수행하고 있음을 알 수 있다. 분류표에 열거된 개념들은 시소러스에서도 반복되고 있고 따라서 자모순 시소러스는 분류표를 참고하지 않고도 그 자체로서 하나의 개체로 사용될 수 있다.



<그림 8-6> LEC 분류표와 색인/시소러스 엔트리 사례

9. 온톨로지

9.1 온톨로지의 정의
9.2 온톨로지의 목적
9.3 온톨로지의 개발 필요성
9.4 온톨로지의 구성요소
9.5 온톨로지의 설계

9. 온톨로지(노상규, 박진수 2007, 9-10)

9.1 온톨로지의 정의

온톨로지의 기원은 철학에서 나온 개념으로 '존재론' 또는 '존재학'이라고 불린다. 온톨로지란 철학의 한 분류로서, 존재하는 것의 형태 또는 본질에 관해 연구하는 학문이다. 좀더 구체적으로 말하자면, 온톨로지란 사물의 기본적인 범주나 세상을 구성하는 구성 요소들을 상징하는 일반적인 개념을 다루는 학문이다. 즉, 온톨로지란 존재하는 것과 그것의 기본적인 범주를 연구하는 학문이라고 할 수 있다. 물론 철학자들마다 존재의 기본적인 범주가 무엇인가에 대한 의견은 다르다. 하지만 많은 철학자들이 물리적 객체(object)들은 구체적인 실재인 반면에 클래스나 클래스의 속성 및 관계는 추상적인 실재라고 말한다(노상규, 박진수 2007, 11)

오늘날 온톨로지는 주로 존재론의 학문적 연구결과물을 컴퓨터 분야에서 차용하여 사용하는 것으로 이해할 수 있으며, 데이터베이스 분야, 시맨틱 웹 분야, 인공지능 분야 등에서 개발이 이루어지고 있다. 이와 같은 분야에서 온톨로지의 정의로 가장 널리 알려진 것은 토마스 그루버에 의한 것으로 그의 정의는 간략할 뿐 아니라 온톨로지의 필요 조건을 잘 표현하였다. 그루버에 의하면 온톨로지란 '공유하는 개념화의 형식적이고 명확한 명세'이다. 이 정의를 풀어서 해석하면 다음과 같다(노상규, 박진수 2007, 21);

• '공유'란 말의 의미는 온톨로지가 '합의된 지식'을 표현해야 한다는 것이다. 여기서 합의된 지식이란 몇몇 개인이 임의로 정한 것이 아니라 관련된 모든 구성원의 동의에 의해 수용되는 개념과 개념들 간의 관계를 표현한 지식을 말한다.

• "개념화'란 특정 영역 또는 분야의 현실 세계와 관련된 개념을 나타내는 추상 모델을 일컫는다.

• '형식적'이란 온톨로지의 내용을 컴퓨터가 읽을 수 있고 처리가 가능한 형태로 표현해야 한다는 뜻이다. 물론 형식성의 정도는 차이가 있을 수 있다.

• '명확한'의 뜻은 특정 영역을 모델링할 때 사용하는 개념들과 이러한 개념들을 사용할 때 적용되는 제약 조건들을 명시적으로 정의해야 한다는 것이다.

결국 온톨로지란 특정 분야의 현실 세계를 모델링할 때 이와 관련된 모든 개인이나 집단들이 합의하여 도출한 개념들을 명시적으로 정의할 뿐만 아니라 컴퓨터가 이해하고 처리할 수 있는 형태로 표현하여 나타낸 용어들의 논리적 집합이다. 이런 관점에서 온톨로지는 세상의 특정 분야에 관련된 용어(개념)들을 정의하고, 이들

간의 관계들로 구성된 일종의 사전인 것이다.

그러나 온톨로지는 단순히 특정 분야를 표현하는 개념들의 의미만을 정의한 것이 아니라 각 개념이 지닌 고유한 속성, 개념들 간의 관계 및 이들 사이의 제약조건, 지식 추론을 위한 공리 규칙과 각 개념의 인스턴스(개체)를 총체적으로 정의함으로써 그 분야의 지식 체계를 컴퓨터가 해석하고 이해하여 처리할 수 있도록 형식화한 표준 명세서이다. 따라서 온톨로지의 궁극적인 목적은 컴퓨터가 해석 · 이해 · 처리할 수 있는 특정 영역의 지식 체계를 모델링하는 것이라고 할 수 있다(노상규, 박진수 2007, 22).

한편 언어학 분야에서의 온톨로지에 대한 정의는 '실세계(혹은 특정 도메인)에 존재하는 모든 개념들과 그 개념들의 속성, 그리고 개념들이 상호간 의미적으로 어떻게 연결되어 있는가에 대한 정보를 가지고 있는 지식 베이스'라고 할 수 있다. 온톨로지가 일반적으로 가지고 있어야 하는 특징들을 정리하면 1) 일정한 체계에 의한 어휘 사전이나 용어의 확보, 2) 특정 영역(specified domain)뿐만 아니라 보편 영역(generic domain)의 기본 개념에 대한 정의와 그들 간의 관계에 대한 명세화, 3) 개념, 관계, 속성 들의 유기적인 집합, 4) 전산적 처리가 가능한 구조화와 구체성, 5) 공유와 재사용의 가능, 6) 논리적 추론, 7) 통합 등을 들 수 있다(한유석, 설근수 2004, 168-169).

온톨로지의 종류는 일반적으로 개념화의 주제 또는 구축 범위에 따라 일반 온톨로지, 특정 영역 온톨로지 등으로 나뉜다. 일반 온톨로지는 공통적인 상위 개념을 표현하여 온톨로지의 재사용에 유용한 온톨로지를 말한다. 또한 자연 언어 표현의 분석과 온톨로지 생성에 중요한 역할을 한다. 특정 영역 온톨로지는 특정 영역에서 이루어지는 업무에 대한 특성화 또는 일반화된 온톨로지라 할 수 있다. 이들 온톨로지는 아직까지 정확하게 구분하기보다는 혼재되어 사용되는 경우가 많다(한유석, 설근수 2004, 169). 이밖에도 구축 대상에 따라 메타데이터 온톨로지, 웹온톨로지, 표현 온톨로지, 특정 업무 온톨로지 등의 종류로 구분될 수 있다.

최근에 온톨로지의 개발 영역이 인공 지능 연구 분야에서 분야별 전문가의 연구 영역으로 확대되고 있다. 온톨로지는 웹에서 자주 접할 수 있는데 웹에 존재하는 온톨로지의 범위는 웹사이트를 범주화한 대규모 텍사노미에서부터 상품과 상품 특성에 대한 범주 정보를 포함한다. 현재 여러 학문 분야에서 그 분야의 전문가들이 정보를 공유하고 해석하는데 활용할 수 있는 표준 온톨로지를 개발하고 있다. 그 예로 의학 분야에서는 SNOMED나 통합의학언어시스템(UMLS)과 같이 표준화 및 구조화된 대규모 어휘 체계를 개발하였다.

9.2 온톨로지의 목적

9.2.1 시멘틱 상호운용성

상호운용성(interoperability)이란 상이란 정보 시스템들이 각각의 고유한 자율성과 다양성을 유지하면서도 마치 하나의 시스템처럼 운용되는 것을 의미한다. 상호운용성은 두 종류로 나눌 수 있는데 첫째, 신택틱 상호운용성(syntactic interoperability)은 XML 기반의 웹 서비스에서 사용되는 표준화된 프로토콜을 사용하여 상이한 소프트웨어 컴포넌트간에 메시지를 주고 받음으로써 시스템간의 상호운용성을 제공한다. 둘째, 시맨틱 상호운용성(sementic interoperability)은 정보 자체에 구체적으로 나타나 있지 않는 암시적 의미나 내재하는 규칙까지도 상호 이해하여 이종 시스템간에 정보의 의미까지도 공유할 수 있는 능력을 말한다.

대부분의 경우 이종 시스템 간에 실시간으로 정보를 공유해야 하는 경우 각 시스템에 저장되어 있는 정보의 의미나 논리적 구조의 이질성으로 인해 기존 시스템을 수정해야 하는 경우가 많으며, 이는 개별 시스템의 독립성과 자율성을 저해한다.

이와 같은 문제를 해결하기 위해 지금까지 약 30년 동안 데이터베이스, 인공지능, 언어학, 기호학 등 여러 학문 분야에서 연구가 활발히 진행되어 왔지만 불행히도 시스템 간의 완벽한 시맨틱 상호운용성을 자동적으로 제공해 주는 방법을 제시하지는 못하고 있다. 그러나 최근에는 많은 학자들이 시맨틱 상호운용성을 제공해 줄 수 있는 핵심 기술을 온톨로지라 보고 이에 관한 연구가 활발히 진행되고 있다. 온톨로지를 이용해 데이터의 의미를 기술할 수 있기 때문에 온톨로지를 사용하여 다양한 형태의 의미 충돌을 해결할 수 있다. 특히 관련 집단의 구성원들 사이에서 합의하여 도출된 개념화를 통해 구축한 온톨로지를 이종 시스템 간의 상호 이해와 시맨틱 조정을 위한 기반으로 사용한다면, 시스템 간의 시맨틱 이질성을 해결하고 시맨틱 상호운용성을 제공하는 매우 중요한 기술이 될 수 있다.

9.2.2 표준화

온톨로지는 특정 영역의 개념 구조를 합의된 지식으로 표현한 것이다. 특정 영역에 사용되는 개념을 표현하는 단어들과 그것들의 관계를 계층적 구조로 나타냄으로써 구성원 모두가 사회적 합의하에 사용할 수 있는 일종의 표준 명세이다. 따라서 온톨로지는 사실상 표준으로서 개발되고 사용될 수 있기 때문에 구성원들 간의 지

식의 공유를 가능하게 해 준다. 뿐만 아니라 표준화는 시맨틱 상호운용성과도 깊게 연관되어 있다. 온톨로지는 표준명세로서의 역할을 수행하기 때문에 응용 프로그램 사이의 정보 및 지식의 공유를 수월하게 한다. 즉 표준화된 온톨로지는 특정 영역의 지식을 문서화하고 재사용할 수 있는 기능을 제공하는 장점이 있다.

9.2.3 커뮤니케이션

온톨로지는 개념적으로나 용어적으로 혼돈을 줄 수 있는 부분들을 단일화된 구조로 명시함으로써 서로 다른 견해나 생각을 가진 구성원들 사이에 공유된 이해와 커뮤니케이션을 촉진한다.

온톨로지는 응용 프로그램이나 소프트웨어 에이전트들이 특정 분야의 개념들이 지닌 의미를 정확히 이해할 수 있도록 하여 컴퓨터 간의 커뮤니케이션을 가능하게 한다. 특히 온톨로지는 어떤 목적을 달성할 수 있는 문제 해결 방법을 일련의 규칙과 제약 조건으로 표현할 수 있기 때문에 인간의 협상과 거래에 대한 지식을 소프트웨어 에이전트에게 제공해서 인간을 대신해서 협상을 효과적이고 지능적으로 수행할 수 있게 한다.

9.2.4 지식 관리와 검색

인터넷 등 IT의 발달로 인해 방대한 양의 정보에 용이하게 접근하는 것이 가능하게 되었지만 원하는 정보를 정확하게 검색하는 것이 갈수록 어려워지고 있다. 뿐만 아니라 정보의 내용 및 형식이 매우 다양해지고 있어 이를 체계적으로 저장하고 관리하며 필요한 지식을 추출하는 것을 더욱 어렵게 하고 있다.

따라서 짧은 시간 내에 오직 자신이 필요로 하는 지식을 찾는 것이 매우 중요하게 되었고 근본적인 문제는 관련 정보를 어떻게 효율적으로 처리하느냐가 아니라 어떤 정보가 관련이 있고 그 정보가 어디에 있는지를 정확히 찾아내는 것이라고 할 수 있다.

온톨로지는 광대한 정보 공간 속에서 우리의 지식관리 및 검색능력을 향상시켜 줄 수 있다. 온톨로지를 이용하여 지식을 검색할 경우 특정 용어와 관련된 다른 지식의 검색도 가능하게 해 준다. 뿐만 아니라 온톨로지를 이용하여 단순한 검색어 기반의 매칭 기술(keyword matching)이 아닌 보다 지능적인 시맨틱 기반의 검색과 필터링 기술로 각 개인이 사용하는 용어의 차이에 관계없이 정확히 필요한 정보만을 검색하는 것이 가능하다. 만일 검색 엔진과 사용자가 온톨로지의 도움을 받는다

면 사용자들은 동일한 검색어지만 다른 의미를 지닌 정보, 즉 관련이 없는 정보를 검색하는 경우들을 피할 수 있을 것이다.

9.3 온톨로지의 개발 필요성

온톨로지는 누가 개발하려 하는가? 온톨로지의 개발 필요성은 다음과 같다(Noy and Deborah 2008):

• 사람과 에이전트 프로그램의 정보 구조에 대한 이해 공유: 어떤 주제에 관한 각 웹 사이트에서 사용하는 모든 용어를 동일한 온톨로지에서 가져온다면, 에이전트 프로그램을 이용해서 서로 다른 여러 사이트에서 추출한 정보를 통합할 수 있다. 그리고 그 에이전트 프로그램은 통합한 정보를 이용하여 이용자 질의에 응답하거나 다른 응용프로그램에 데이터를 제공할 수 있다.

• 특정 분야의 지식 재활용: 특정 분야의 지식을 재활용하는 것은 최근에 각광받는 온톨로지 연구 분야이다. 예를 들어 여러 분야의 온톨로지 모델은 각각 시간 개념을 필요로 한다. 시간의 표현에는 시간차, 시점, 시간의 상대 측정 등에 대한 개념을 포함한다. 한 분야의 연구자 집단이 이러한 온톨로지를 상세히 개발한다면 다른 연구자 집단은 그것을 용이하게 자기 분야에 재활용할 수 있다. 뿐만 아니라 대규모의 온톨로지를 구축할 때에 대규모 연구 영역의 각 부분을 구성하는 여러 기존 온톨로지를 통합할 수 있다.

• 해당 분야의 가설을 명시: 온톨로지 구현에 밑바탕이 되는 해당 분야의 가설(assumptions)을 명시적으로 나타내면, 해당 분야에 대한 우리의 지식이 변할 때 관련 가설도 쉽게 수정할 수 있다. 또한 해당 분야의 지식에 대한 명확한 명세는 그 분야에서 사용하는 용어 의미를 학습해야 하는 신규 이용자에게도 유용하다.

• 운영상의 지식에서 해당 분야의 지식을 분리: 운영상의 지식에서 해당 분야의 지식을 분리하는 것 또한 온톨로지의 공통적인 용도 중 다른 하나이다. 필수 명세(설명서)에 따라 부품에서 상품을 조립하는 업무를 기술하고, 상품이나 부품과 독립적으로 조립 프로그램을 구현할 수 있다. 그리고 PC-부품과 특성에 대한 온톨로지를 개발하고, 이것을 주문 제작용 PC를 조집하는 알고리즘에 적용할 수 있다. 우리가 엘리베이터 부품 온톨로지에 동일한 알고리즘을 적용시키면, 엘리베이터를 조립할 수도 있을 것이다.

• 해당 분야의 지식 분석: 일단 용어에 대해 선언적인 명세를 작성하면, 해당 분야 지식을 분석할 수 있다. 용어에 대한 형식 분석은 기존 온톨로지를 재활용하거

나 확장할 때 유용하다.

9.4 온톨로지의 구성요소

온톨로지로 분류되기 위해서는 적어도 용어와 각 용어의 내포를 명확히 명시할 수 있는 어떤 형식을 포함해야 한다. 즉 온톨로지는 개념을 표현하는 용어에 관한 정의와 이러한 용어들이 어떻게 상호 관계를 가지고 있는지를 명시함으로써 모델링하고자 하는 세계에 대한 체계적이고 논리적인 틀을 제공해야 한다.

대부분의 온톨로지는 다음과 같이 '개념(concept)', '속성(property)', '관계(relationship)', '제약조건(constraint)', '공리(axiom)', '인스턴스(instance)'의 여섯 가지 구성요소로 이루어져 있다. 온톨로지가 반드시 이들 구성요소를 갖추어야만 성립되는 것은 아니지만 추론 능력을 포함한 다양한 기능을 제공하기 위해서는 가능한 위에서 언급한 모든 구성요소를 포함하는 것이 바람직하다(노상규, 박진수 2007, 2-26).

• 개념(concept): 개념이란 현실 세계에서 존재하는 것에 대한 일반적이고 본질적인 인식이나 지식을 말한다. 개념에는 오감을 통해 물리적으로 느낄 수 있는 것뿐만 아니라 정신세계에서만 존재하는 추상적인 것을 모두 포함한다. 개념은 '사물(thing)'을 표현하는 단위로서 '단순 개념(elementary concept)'과 '합성 개념(composite concept)'의 두 가지로 나뉜다. 단순 개념은 더 이상 나누어지지 않는 기본 개념을 말하고, 합성 개념은 어떤 형태로든 연관성을 지니고 있는 단순 개념들이 결합하여 이루어진 개념이다. 예를 들어 '컴퓨터'는 'CPU', '메모리', '하드디스크' 등 다수의 개념과 부분(part-of) 관계로 성립된 합성 개념이다.

• 속성(property): 속성은 개념에 근본적으로 속해 있는 성질을 말한다.

• 관계(relationship): 관계란 개념들 사이의 상관 관계를 말한다. 개념은 주로 관계를 통해 다른 개념들과 연결되어 있다. 대표적인 관계의 유형으로는 상위 개념과 하위 개념 간의 계층 구조(hierarchical structure)를 형성하는 상속 관계(inheritance relationship)가 있다. 일반적으로 '~이다(is-a)' 또는 '~의 종류이다(is-a-kind-of)'로 표현된다. 상속 개념 이외의 개념 간의 계층 구조를 형성하는 관계로는 합성 관계가 있으며 부분-전체 관계(part-whole relationship)가 가장 널리 알려져있다.

• 제약조건(constraint): 제약조건이란 개념들 간의 관계나 속성의 값에 관한 제한 규정을 말한다. 예를 들어 컴퓨터와 CPU의 관계를 정의할 때 컴퓨터는 반드시

1개 이상의 CPU를 가져야 한다는 제한 규정을 정의할 수 있다. 이러한 제약 조건은 지식을 표현하거나 추론을 할 때 유용하게 사용된다.

• 공리(axiom): 추론의 기본이 되는 명제로서 증명을 할 수 없거나 증명을 요하지 않는 '참(true)'으로 인정되는 문장을 말한다. 모든 공리는 제약조건이며 논리적으로 정확성을 검증하거나 새로운 사실을 증명할 때 유용하게 사용된다. 제약조건과 공리는 주로 일차 논리나 이차 논리를 사용해서 표현된다.

• 인스턴스(instance): 인스턴스는 '개체'라고도 하며, 각 개념의 실례를 말한다.

9.5 온톨로지의 설계(Noy and Deborah 2008)

9.5.1 대상 분야와 범위 결정

온톨로지 개발을 위해 온톨로지에서 다루는 대상 분야와 용도는 무엇인가에 대한 답이 필요하다. 또한 온톨로지를 통해 어떤 유형의 정보 질의에 대한 응답을 제공할 것인지 그리고 온톨로지 이용자와 관리자는 누구인가를 결정해야 한다.

온톨로지의 범위를 결정하는 방법 중 하나는 온톨로지를 이용해 구축한 지식베이스가 응답해야 하는 질문 목록인 적합성 질문(competency questions)을 작성하는 것이다. 온톨로지가 이와 같은 유형의 질문에 응답하는데 필요한 정보를 포함하고 있는가? 이 응답이 구체적인 상세성 수준을 요구하는가 아니면 특정 분야의 대표성을 요구하는가? 이 적합성 질문들은 단지 개요만 제공하면 되는 것으로 망라적일 필요는 없다.

9.5.2 기존 온톨로지 검토

특정 분야나 업무에 대하여 누군가가 먼저 해 놓은 것이 있는지, 기존 정보원을 수정하고 확장할 수 있는지의 여부를 확인하는 것은 항상 가치있는 일이다. 개발하려는 시스템이 특정 온톨로지나 통제 어휘를 포함하고 있는 다른 응용프로그램과 상호작용할 필요가 있다면 기존 온톨로지를 재사용하는 것은 필수요건이다. 많은 지식-표현 시스템이 온톨로지를 가져오거나 내보낼 수 있기 때문에 전자 형태로 이용가능하며 온톨로지 개발 환경으로 가져올 수 있다.

재사용할 수 있는 온톨로지 라이브러리의 예로는 Ontolingua 온톨로지 라이브러리(http://www.ksl.stanford.edu/software/ontolingua)나 DAML 온톨로지 라이브러리(http://www.daml.org/ontologies) 등을 들 수 있다.

9.5.3 온톨로지의 주요 용어 열거

이용자에게 설명하고 싶은 모든 용어에 대한 목록을 만드는 것이 유용하다. 우리가 논의하고 싶은 용어는 무엇인가? 그 용어들은 어떤 속성을 가지고 있는가를 검토하는 것이다. 다음의 두 단계-클래스 계층 개발과 개념의 속성 개발은 상당히 얽혀 있어서 어느 것이 먼저이고 어느 것이 나중이라고 구분하기 어렵다. 일반적으로 각 계층에 속한 개념들에 대한 간단한 정의를 작성하고, 각 개념이 가진 속성을 기술해 나간다. 이 두 단계는 온톨로지 설계 과정에서 가장 중요한 단계이다.

온톨로지 구축에 필요한 가장 기초적인 자원은 어휘라 할 수 있다. 이러한 어휘들의 확보 방법은 다양하지만, 일반적으로 일반 어휘나 전문 용어들을 체계적으로 정리한 사전, 백과사전, 전문 용어 사전 등을 이용하거나, 특정 문서나 웹문서에서 어휘를 추출하여 사용한다. 하지만 온톨로지 구축에 이용될 어휘들에 대한 형태적 · 의미적 기준을 확립하기 쉽지 않기 때문에 이러한 기초 자원을 사용함에 있어 주의해야 한다. 즉 기초 자원 설정 및 활용 기준을 어떻게 하느냐에 따라 온톨로지의 기초적인 구축 원리가 확립될 수 있는 것이다(한유석, 설근수 2004, 172)

9.5.4 클래스와 클래스 계층의 정의

클래스는 대부분의 온톨로지의 핵심 요소로, 대상 분야의 개념을 기술한다. 클래스는 상위 클래스보다 더욱 특정한 개념을 표현하는 하위 클래스를 가진다.

클래스의 계층구조 개발에 사용가능한 접근방식으로는 1) 하향식의 경우 대상 분야에 속하는 가장 일반적인 개념에 대한 정의로부터 시작해서 순차적으로 개념을 구체화시킨다. 2) 상향식은 계층구조의 가장 아래 부분에 해당되는 구체적인 개념의 정의에서부터 시작해서 순차적으로 이 개념을 보다 일반적인 개념으로 묶는 방식이다. 3) 조합식은 하향식과 상향식 접근방식을 혼용하는 것이다. 가장 대표적인 개념을 우선 정의한 다음 그 개념을 적절하게 일반화 및 구체화시킨다.

이들 세 가지 방법 중에서 다른 것보다 본질적으로 나은 것은 없다. 조합식 접근방법은 흔히 많은 온톨로지 개발자들에게 가장 용이한 접근방식인데 '중간 수준'이라는 개념이 각 분야에서 가장 기술하기 용이한 개념인 경향이 있기 때문이다. 어떤 접근방식을 선택하든지 대개 클래스를 정의하는 것부터 시작한다.

클래스의 계층 구조는 온톨로지의 용도, 응용프로그램에서 요구하는 상세성 정도, 개인적인 선호도, 여타 다른 모델과의 호환성 요건에 따라 달라진다.

9.5.5 클래스의 속성 정의

클래스만으로는 적합성 질문에 응답할 수 있는 충분한 정보를 확보할 수 없다. 클래스를 정의하고 나면 다음에는 개념들의 내부 구조를 기술해야 한다. 즉, 목록에 있는 각 속성들이 어떤 클래스를 기술하는지 결정해야 한다. 특정 클래스의 모든 하위 클래스는 해당 클래스의 속성을 상속하며, 이 속성들은 클래스에 연결된 슬롯(패싯)이 된다. 예를 들어 와인의 속성은 색깔, 농도, 향, 당도, 생산지 등이 있으며 이들 속성은 와인 온톨로지의 슬롯이 될 것이다. 사물의 속성에는 내재적 속성(예: 와인의 향)과 외재적 속성(예: 와인명, 생산지 등), 다른 객체와의 관계속성 등의 유형이 있다.

9.5.6 슬롯 패싯의 정의

슬롯은 값의 유형, 허용되는 값, 값의 개수 및 해당 슬롯이 가질 수 있는 값의 다른 특성을 기술하는 여러 패싯을 가질 수 있다.

슬롯이 가질 수 있는 값의 수는 단수나 복수를 허용할 수 있으며, 슬롯 값의 유형은 문자열, 숫자, 블리언, 열거형, 인스턴스형 등으로 정의될 수 있다.

9.5.7 개별 사례 생성

마지막 단계는 계층 관계에서 클래스의 개별 사례를 생성하는 것이다. 클래스의 개별 사례를 정의하는 것은 1) 클래스의 선정, 2) 클래스의 개별 사례 생성, 3) 슬롯 값을 채우는 과정을 요구한다.

10. 온톨로지 구축 사례

10. 온톨로지 구축 사례

10.1 개요

지금까지 구축된 온톨로지는 주로 지식 추출, 자연어 검색, 분야별 어휘 사전, 분류체계 용도 등으로 사용되고 있다. 지식추출 용도란 인공 지능(artificial intelligence)과 같은 분야에서 인간의 지식과 추론 엔진을 함께 구축하여 컴퓨터로 하여금 인간과 같은 사고를 할 수 있도록 온톨로지를 구축한 경우를 말한다. 자연어 검색 용도는 웹 또는 응용프로그램에서 일상적으로 우리가 사용하는 문구와 문장을 그대로 입력하여 우리가 원하는 결과를 얻을 수 있도록 흔히 사용되는 구문 정보 등을 온톨로지에 담아 사용하는 것을 의미한다. 어휘 사전 용도란 온톨로지를 사전으로 구축하여 특수 분야에서 사용하는 용어와 그 용어들 간의 관계를 저장해 놓은 것으로, 가장 많이 사용되는 온톨로지의 한 종류이다. 마지막으로 분류 체계 용도란 디렉터리 혹은 상품 분류와 같이 내용들이 서로 계층적인 관계를 가지고 있을 때, 그들 간의 관계를 표현할 수 있도록 온톨로지를 사용하는 것을 의미한다(노상규, 박진수, 2007).

10.2 워드넷(WordNet)[6]

WordNet은 어휘 지식에 대한 인지 언어학의 연구 성과를 토대로 미국의 프린스턴대학이 구축해 온 어휘 데이터베이스이다. 단어를 중심으로 접근하는 일반 사전과는 달리 보다 효율적으로 검색이 이루어지도록 개념을 통한 접근을 시도한 것으로, 초반에는 언어심리학을 위한 사전으로 개발되었지만 점차 규모가 커져 일반 사전으로까지 그 범위가 확장되었다.

WordNet의 목적은 크게 두 가지로 나누어 볼 수 있다. 첫 번째 목적은 의미를 중심으로 접근할 수 있는 일반 사전의 기능과 동의어, 반의어를 제시할 수 있는 시소러스의 조합을 만들어 내는 것이다. 두 번째 목적은 자동 텍스트 분석과 인공 지능을 지원하는 것이다. 이것은 일반 이용자가 보다 편하게 웹 사이트를 이용할 수 있게 해준다.

WordNet은 표면적으로는 시소러스와 유사하지만 몇 가지 차이점이 있다. 첫째, 단어의 형태 뿐 아니라 단어의 특정 의미를 연결해주기 때문에 그 결과 네트워크

6) http://wordnet.princeton.edu/

내에서 의미상 비슷한 뜻을 갖는 근접한 다른 단어를 찾을 수 있게 해준다. 둘째, 시소러스 내에서의 단어들의 그룹핑은 의미의 유사성 이외에 다른 명확한 패턴을 보여주지 못하는 반면, 단어들 간의 의미관계를 표시하고 있다. 이런 점에서 WordNet은 전통적인 사전이나 시소러스라고 하기 어렵지만 이들의 장점을 결합한 어휘집이라고 할 수 있다.

WordNet에서 단어의 기본적인 의미 관계는 동의 관계로서 동의어 집합을 synset이라 하고 이를 기본 구조로 한다. 이 synset은 표현하는 의미가 무엇인지 명시하지 않으며 다만 그와 같은 의미가 존재하고 있다는 사실만 나타낸다. WordNet에서 개념은 그 개념을 표현하기 위해 사용되는 동의어의 집합으로서 다의성과 동의 관계를 이용하여 의미를 최대한 정확히 표현하고자 한다. 그래서 단어와 그 의미와의 관계는 단어형-의미 행렬을 전제로 한다. 예를 들어 {board, plank}와 {board, committee}는 각각 하나의 의미를 표현하는 동의어 집합이다. 이 집합을 통해서 <표 10-1>와 같이 'board'라는 다의어의 여러 의미가 구체화된다.

<표 10-1> 'board'와 관련된 단어-의미 행렬

의미	board	plank	committee	card	…
{board, plank}	○	○			
{board, committee}	○		○		
{board, card}	○			○	
:					:

synset을 정확히 구현하기 위해 개념 간의 관계 유형은 다음과 같다.

- 동의 관계: synset을 구성하는 기본 관계로서 동사나, 형용사, 부사에는 엄밀한 의미의 동의어가 많지 않으며 기준의 강도에 따라 차이가 심하다.
- 반의 관계: 명사 간에도 존재하지만 특히 형용사와 부사에서 중요한 관계이다.
- 하의 관계: 상의 관계(?)와 함께 명사 synset 간의 계층 관계를 표현한다. 시소러스에서의 계층 관계와 유사하다.
- 부분 관계: 전체 관계와 함께 부분-전체 관계를 표현한다. 시소러스에서 BTP/NTP로 표현하는 부분-전체 관계와 동일하다.
- 함의 관계: 동사간의 내포 관계나 인과 관계를 표현하는 데 쓰인다.
- 양식 관계: 함의의 일종이지만 WordNet에서는 범주를 달리 설정하고 있다.

<표 10-2> WordNet과 시소러스의 개념 간 관계 유형 비교

WordNet 관계 표현		시소러스의 관계 표현
(synonymy)		등가관계와 거의 같은
반의(antonymy)	△	연관관계에 포함됨
상의/하의(hypernymy/hyponymy)	○	계층관계와 유사함
부분/전체(meronymy/holonymy)	○	부분-전체관계(BTP/NTP)와 같음
양식(troponomy)	×	없음, 계층관계와 비슷함
함의(entailment)	×	없음, 연관관계와 비슷함

이상의 관계 유형은 기존의 시소러스가 채택하고 있는 계층 관계, 등가 관계, 연관 관계보다는 다양하지만 실제로 <표 10-2>와 같이 큰 차이가 없음을 알 수 있다. 특히 명사에 적용되는 관계 유형만 보면 반의 관계 이외에는 시소러스와 WordNet의 관계가 거의 동일함을 알 수 있다. 시소러스에서 명사 간의 계층 관계를 세분한 속종 관계, 사례 관계는 WordNet에서 상의/하의 관계로 표현되지만 실제로 이 두 관계를 구분하는 시소러스는 많지 않다. 반의 관계는 시소러스에서는 일반적으로 연관 관계로 취급되고 있다. 양식 관계나 함의 관계는 주로 동사, 형용사 등의 용언에 적용되기 때문에 명사를 주로 취급하는 시소러스에서는 규정하지 않고 있다.

WordNet은 품사별로 구축되었기 때문에 개념 분류 체계도 품사별로 이루어져 있으며, 그 중 명사는 25개 범주와 12단계까지 계층 수준으로 전개되어 있다. 각 계층의 최상위 명사 개념은 <표 10-3>과 같다.

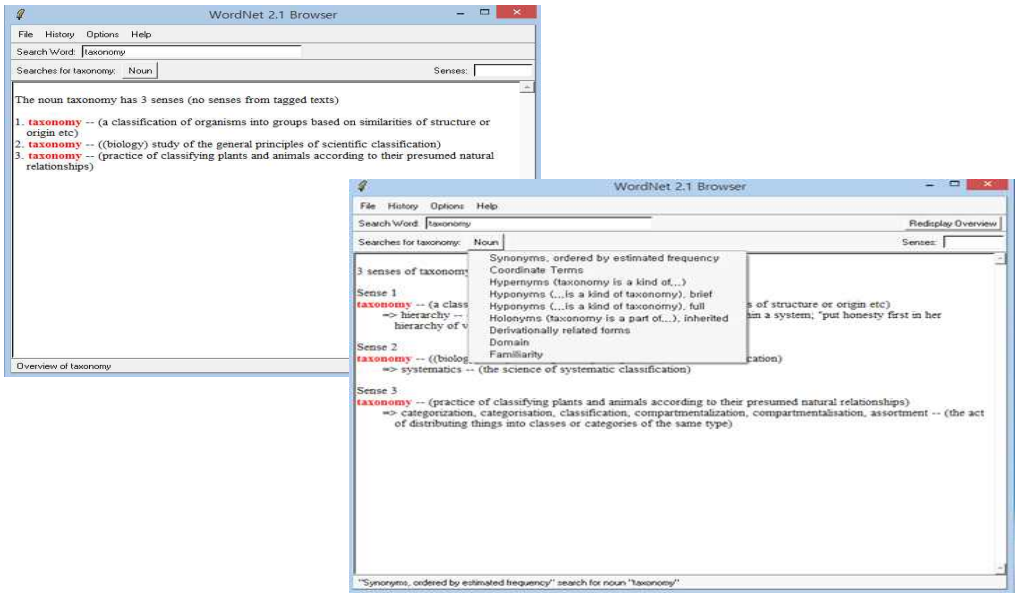
현재 WordNet은 웹 사이트에서 온라인 검색으로 사용할 수 있는 방법과 프로그램을 다운로드 받아서 설치한 수 사용할 수 있는 방법이 있다. WordNet의 검색은 일반 사전을 이용했을 때와 같이 단어의 사전전 정의를 살펴볼 수 있으며, 나아가 인지적 동의어들의 집합으로 분류하고 이 집합들을 의미 관계로 엮어 놓았기 때문에 검색 단어의 자주 사용되는 동의어(synonyms), 동의어의 더 구체적인 동의어(coordinate terms), 해당 단어의 일부 혹은 전체가 포함된 다른 단어의 정의(hypernyms), 해당 단어와 관련된 다른 형태의 단어, 그리고 해당 단어가 사용되는 일반적인 문장 구조 등과 같은 정보를 얻을 수 있다.

예를 들어, <그림 10-1>은 다운로드한 프로그램을 통해 'taxonomy'라는 단어를 검색한 결과를 보여주는 화면이다. 검색결과 3개의 의미를 가진 다의어로 사용되고 있음을 알 수 있으며, 'Noun'이라는 메뉴 아래의 'Synonyms, ordered by estimated frequency'를 선택하여 동의어를 찾아볼 수 있다. 3번째에 정의되고 있

는 'taxonomy'의 의미는 '추정되는 자연적 관계에 따라 동식물을 분류하는 일'이며 이것의 동의어로 '사물을 동일한 유형의 주류나 범주로 배분하는 행위'의 의미를 지닌 'catecategorization, categorisation, classification, compartmentalization, compartmentalisation, assortment'의 단어들이 사용되는 것을 알 수 있다.

<표 10-3> WordNet의 명사 최상위 계층

명사 synset	계층 내용(synset)	synset의 예
{act, action, activity}	명사	{accomplishment, deed}
{animal, fauna}	동물 명사	{game, prey, quarry}
{artifact}	인공물 명사	{cloth, babric, textile} / {device}
{attribute, property}	속성 명사	{age} / {power}
{body, corpus}	신체부위 명사	{body parts}
{cognition, knowledge}	인지 명사	{intellect, mind}
{communication}	커뮤니케이션 명사	{language, linguistic communication}
{event, happening}	사건 명사	{human event} / {social event)
{feeling, emotion}	감정 명사	{mood} / {excitement}
{food}	음식 명사	{pizza, pizza pie}
{group, collection}	집합 명사	{team, squad}
{location, place}	위치 명사	{point, spot} / {region, area}
{motive}	동기 명사	{motive, need}
{natural object}	자연물 명사	{cloud} / {cosmos, universe}
{natural phenomenon}	자연현상 명사	{luck, fortune, chance, hazard}
{person, human being}	인물 명사	{Adam, Robert Adam}
{plant, flora}	식물 명사	{clover, trefoil}
{possession}	소유 명사	{asset} /{liability}
{process}	과정 명사	{natural process} / {increase}
{quantity, amount}	수량 명사	{definite quantity} / {relative quantity}
{relation}	관계 명사	{difference} / {disjunction}
{shape}	형상 명사	{shape, form}
{state, condition}	상태 명사	{panic, scare}
{substance}	물질 명사	{solid} / {liquid}
{time}	시간 명사	{daytime / time of day}



<그림 10-1> WordNet에서의 'taxonomy' 검색 결과 화면

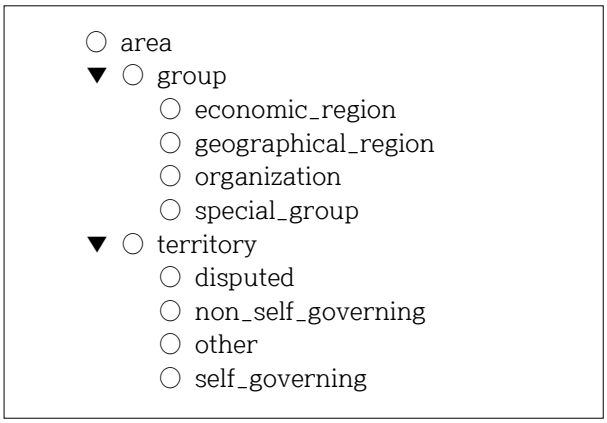
10.3 지정학 온톨로지(Geopolitical Ontology)[7]

2002년에 UN 산하 FAO(Food and Agriculture Organization)는 FAO 테마에 따라 국가 프로필에 접근할 수 있는 혁신적인 정보 검색 시스템을 개발하였다. 2006년부터 해당 데이터를 FAO 지정학 온톨로지를 기반으로 구축하여 2008년에는 베타 버전이, 2010년에는 1.1 버전을 배포하였으며 국제 사회에서의 자원 재활용을 위해 온톨로지를 개방하였다. 그러므로 온톨로지는 상호운용성 및 표준 데이터 공유의 편리성에 주안을 두고 정보 관리를 향상시키는 것을 목표로 개발되었다. 온톨로지의 참조 데이터를 쉽게 활용할 수 있도록 다양한 웹 서비스와 온톨로지 모듈 메이커(ontology module maker)를 제공하고 있다.

따라서 FAO 지정학 온톨로지는 데이터 소스 및 구조가 공개되어 있으므로 온톨로지 구축 과정을 용이하게 이해하고 참조할 수 있는 좋은 사례가 될 수 있다. 대부분의 온톨로지 모델링이 하향식(Top-down)으로 이루어지는 반면에 지정학 온톨로지는 상향식(bottom-up)에 의한 접근방식으로 이루어진 것도 특징이 된다. 지정학 도메인과 관련하여 실제 객체(objects)가 되는 국가, 비자치 지역(non-self-governing territories), 지리적 지역(geographic regions), 경제적 그룹(economic groups) 등은 고유의 국가명을 가진 인스턴스들로부터 수집되고 시행되었다. 인스턴스들은 공통되는 특성에 따라 8개의 특정 클래스-자치지역, 비자치 지

7) http://www.fao.org/countryprofiles/geoinfo/en/

역, 분쟁지역, 기타지역, 경제적 지역, 지리적 지역, 기관/조직, 특수 집단-로 범주화되었다. 지역을 구분하는 세 개의 일반적인 용어(area, group, territory)는 계층구조를 만드는 클래스로 사용되고 있다(<그림 10-2> 참조).



<그림 10-2> 지정학 온톨로지의 클래스 계층구조

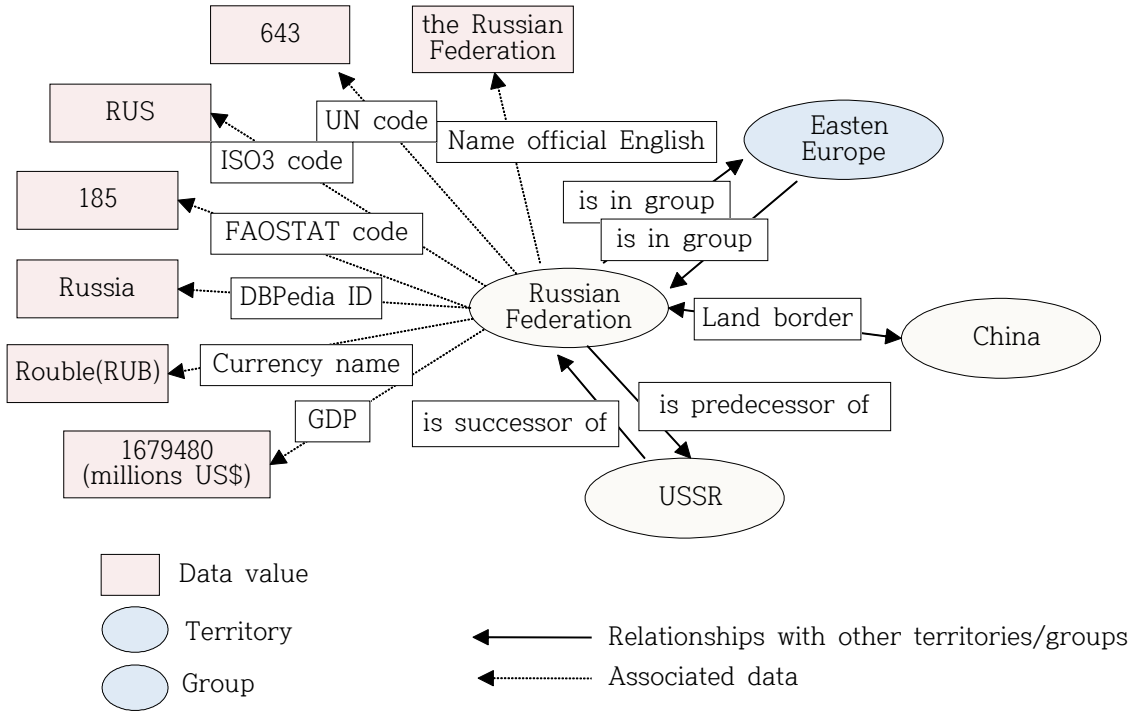
또한 지정학 및 지리적 개체들 간에 사용될 수 있는 관계 유형의 예는 <표 10-4>와 같으며, 클래스와 속성 간의 관계 통해 지정학 온톨로지는 100개 이상의 공리(axioms)를 수행할 수 있다. 이밖에도 온톨로지를 구축하기 위한 국제 코딩 시스템의 사용, 다국어 표현, 데이터 소스의 구조적 기술, OWL version, RDF 변환 등을 공개하고 있다.

<표 10-4> 지정학 온톨로지의 관계 사례

	역관계명	관계 설명	관계(예)
hasBoderWith	hasBoderWith	인접국간의 지리적 관계	Spain hasBoderWith Portugal Portugal asBoderWith Spain
isInGroup	hasMember	집단과 회원국 관계	France isInGroup Europe Europe hasMember France
isPredecessorOf	isSuccessorOf	지명의 역사적 변화	Russian Federation isSuccessorOf USSR USSR isPredecessorOf Russian Federation

지정학 온톨로지의 활용은 UN 산하 기관, 정부 기관, 의사결정자, 연구자 등 국가-기반의 정보를 다루는 이용자에게 참고 자료를 지원해주는 특별한 자원이 될 수 있다. 특별히 지정학 모듈 메이커(Geopolitical Ontology module maker)는 다양

한 포맷으로 온톨로지를 모듈별로 추출하여 용이하게 다른 정보 시스템에 적용가능하게 해주며, 온톨로지 웹 서비스는 이용자들의 국가 기반 온톨로지의 탐색을 도와준다. <그림 10-3>은 '러시아'를 예로 들어 표현되는 온톨로지를 보여 준다.



<그림 10-3> '러시아' 지정학 온톨로지 사례

10.4 사이크(Cyc)[8)]

사이크는 컴퓨터로부터 새로운 지식의 추론이 가능하도록 도와주는 인공 지능의 지식 추출용 온톨로지라고 할 수 있다. 1984년도에 '생각하는 기계'를 만드는 사이크 프로젝트가 착수되어, 1994년도에 MCC(Microelectronics and Computer Technology Company) 연구 컨소시엄의 확장된 형태로 창립된 Cycorp에 의해 지금까지 인공 지능 분야의 대표적인 연구로 진행되고 있으며, 2001년 오픈사이크(OpenCyc)와 2005년 리서치사이크(ResearchDyc)가 웹사이트에 공개되면서 온톨로지의 내부 구조가 알려지게 되었다(<그림 10-4> 참조).

사이크 프로젝트는 사이크 시스템이 사용자와 자연어로 의사소통하고, 기계 학습(machine learning)을 통해서 자체적으로 지식을 창출해 낼 수 있으며, 인간이 가지고 있는 일반 상식과 전문 지식을 모두 포함하고 있는 지식 베이스를 구축하는

8) http://www.cyc.com/

것을 목표로 한다. 따라서 사이크 프로젝트는 지식 베이스의 일부를 온톨로지로 구축하여 컴퓨터가 이해할 수 있도록 연구하는 것이다.

<그림 10-4> OpenCyc와 ResearchCyc 웹페이지 화면

사이크의 전체적인 구조는 다음과 같이 지식 베이스, 추론 엔진, 월드, 인터페이스의 네 부분으로 이루어진다.

- 지식 베이스: 사이크의 핵심이라고 할 수 있으며, 50만 개 이상의 개념(concept)과 개념을 정의하는 2만 6천개 이상의 관계를 사용하여 5백만 개 이상의 선언적 사실(assertion)을 저장하고 있는 시스템의 온톨로지이다. 이곳에 인간의 지식을 담고 있고 컴퓨터가 이곳에서 지식을 추출한다. 이곳에 저장된 지식은 일상적인 상식부터 화학, 생물학, 군사 등의 전문가가 소유하는 지식을 모두 포함하고 있다. 또한 어휘와 문법적인 내용을 포함하여 사람과 컴퓨터가 자연어를 통해서 의사소통할 수 있도록 돕는다.
- 추론 엔진: 지식 베이스에 저장되어 있는 선언적 사실과 함께 외부의 다른 시스템(데이터베이스, 웹사이트 등)으로부터 얻는 정보를 이용하여 결론에 도달하거나 새로운 추측을 할 수 있는 능력을 가진다.
- 월드(world): 이전에 사용했던 선언적 사실들을 스냅샷의 형태로 저장한 파일이다. 이 파일의 도움으로 새롭게 사이크가 수행될 때, 이전의 내용들이 무결성 확인 없이 즉시 메인 메모리로 로드될 수 있다.
- 사용자 인터페이스: 사이크 브라우저는 사용자가 지식 베이스의 내용을 질의하거나 탐색, 수정할 수 있도록 한다.

사이크 시스템은 입력한 검색어 혹은 검색 문구를 분석한 후에 두 개의 창으로 검색 결과를 출력하는데, 왼쪽 창은 검색 결과를 계층적으로 표현한 전체적인 개요이며, 오른쪽 창은 왼쪽 창에서 선택한 카테고리에 대한 검색결과와 관련된 선언적

사실들을 보여준다.

사이크가 적용된 사례에는 미 국방부의 해커 탐지용 시스템, 검색 효율성을 향상시키기 위한 인터넷 검색 엔진 라이코스(Lycos) 등이 있다.

10.5 오픈 디렉터리 프로젝트(Open Directory Project)[9)]

인터넷의 급속한 성장으로 인해 넘쳐나는 정보 때문에 검색 엔진이나 웹 디렉터리는 더 이상 사용자가 만족할 만한 결과를 제공하지 못하게 되었다. 이러한 이유로 인터넷 자체를 정리할 수 있는 방법이 필요하게 되었고, 그 결과 오픈 디렉터리 프로젝트(ODP, dmoz)가 등장하게 되었다.

ODP는 1988년도 자유 소프트웨어를 지향하며 시작되었으며 웹 디렉터리, 즉 웹 컨텐츠 분류 체계용 온톨로지를 구축하는 것을 목표로 하고 있다. ODP의 가장 큰 특징은 옥스퍼드 영어 사전의 편찬 방법과 유사하게 자발적으로 참여자들에 의해 구축되었다는 것이다. 60만개에 이르는 카테고리로 분류된 400만 개 이상의 웹사이트 디렉터리 정보로 구성된 ODP 온톨로지는 7만 명이 넘는 자발적 참여자들에 의해서 꾸준히 업데이트되고 있다. 가입한 사용자 누구든지 직접 디렉터리를 수정할 수 있으며, 구축된 디렉터리 정보는 누구나 100% 무료로 사용 가능하다.

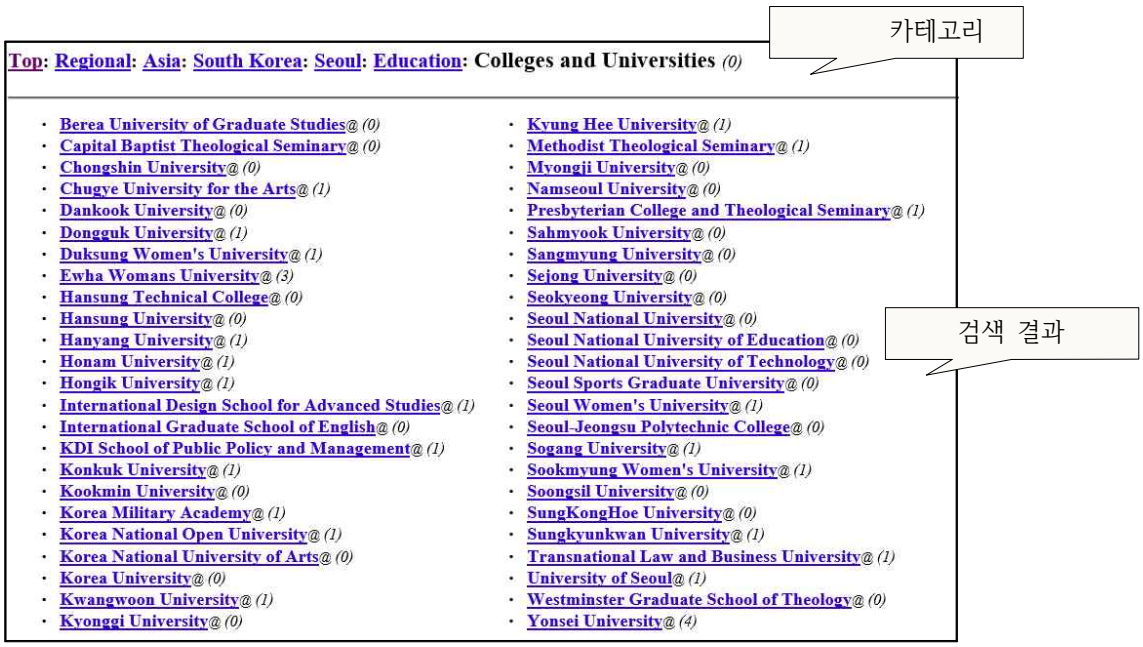
ODP는 실제로 넷스케이프, AOL, 구글, 라이코스(Lycos) 등과 같은 유수의 검색 엔진과 포털 사이트에 디렉터러 서비스를 제공하고 있다. 예전에는 검색 엔진 및 포털 사이트가 자체적으로 개발한 디렉터리를 사용하는 것이 일반적이었지만, 이러한 사이트들이 공통으로 ODP를 도입함으로써 일반 사용자가 해당 사이트를 처음 방문하더라도 익숙한 분류 카테고리를 통하여 용이하게 원하는 사이트에 접근할 수 있게 된다.

<그림 10-5>와 <그림 10-6>은 ODP가 디렉터리 간의 관계를 계층적으로 담고 있는 웹 컨텐츠 분류 체계용 온톨로지로서의 기능을 보여주고 있다. 예를 들어 서울에 소재하는 대학교의 리스트를 원하는 경우 메인 화면에서 'Reference' 아래 하위 메뉴로 포함되어 있는 'Education'을 선택하고 계속하여 계층 구조에 따라 'Colleges and Universities' → 'Asia' → 'South Korea' → 'Seoul'의 하위 메뉴를 순서대로 선택하면 용이하게 원하는 검색 결과를 얻을 수 있다. 디렉터리 서비스는 특정 영역 내의 토픽들이 서로 어떻게 연관되어 있는지를 이해하는 데 도움을 줄 뿐만 아니라 검색에 유용한 용어들을 제안해 주기 때문에 넓은 카테고리로부터 검

9) http://www.dmoz.org/

색을 좁혀나가는데 매우 유용하게 사용될 수 있다.

<그림 10-5> ODP 메인 화면



<그림 10-6> ODP의 검색 사례 화면

10.6 로제타넷(RosettaNet)[10]

로제타넷은 1998년에 설립되었으며 현재까지 계속 확장되고 있고 ebXML과 더불어 전자 거래 문서의 국제 표준으로 폭넓게 사용되고 있다. 로제타넷은 전세계적으

10) https://www.rosettanet.org/RosettaNet

로 정보 기술, 전자 부품, 반도체 분야의 500여 업체로 구성된 하나의 컨소시엄으로서 개발한 기업간 전자 상거래 표준이며, 기업이 직접 적용하고 있기 때문에 실용적이라는 평가를 받고 있다.

기술적 측면에서 로제타넷은 특정 표준 요소나 독점적인 솔루션에 치중하는 다른 분야의 표준화 노력과는 다르게 비즈니스 사전과 기술 사전, 구현 프레임워크, 비즈니스 메시지 관리 체계, 프로세스 정의를 지원하는 개방형 표준을 제공함으로써 완벽한 비즈니스 프로세스 아키텍처를 구축하는 것을 목표로 한다.

로제타넷의 핵심 구성 요소 중에서도 용어 사전은 사전형 온톨로지라고 할 수 있는데 비즈니스 거래를 하는데 필요한 기반으로 여러 회사에서 다양하게 사용되는 용어들을 정의해 놓은 곳이다. 용어 사전은 크게 비즈니스 관련 데이터를 정의하는 비즈니스 사전(RNBD: RosettaNet Business Dictionary)과 전자 부품과 정보 기술 제품들의 속성을 포함하는 기술 사전(RNTD: RosettaNet Technical Dictionary)으로 나누어진다.

로제타넷을 사용하여 업무의 효율성을 높인 사례로 HP는 판매회사, 대리점 사이의 프로세스 운영에 소요되는 시간을 며칠 단위에서 몇 분 단위로 단축했다고 한다. 이는 빠른 속도, 원활한 정보의 흐름, 전송 데이터의 정확성, 불필요한 부가 정보의 유입 차단, 치밀한 계획 및 예측 등이 가능했기 때문이며, 결과적으로 로제타넷이 제공하는 합의된 프로세스(PIP: Partner Interface Process)의 덕택이라고 할 수 있다.

<참고문헌>

고영만 외. 2005. 국가지식정보자원 표준분류체계 연구. 서울: 한국정보문화진흥원.

김태수. 2000. 분류의 이해. 서울: 문헌정보처리연구회.

김포옥, 백항기. 2011. 문헌분류론. 개정판. 고양: 조은글터.

노상규, 박진수. 2007. 인터넷 진화의 열쇠 온톨로지: 웹 2.0에서 3.0으로. 서울: 가즈토이.

문헌정보학용어사전. 2010. 개정판. 서울: 한국도서관협회.

윤희윤. 2013. 정보자료분류론. 개정증보 제4판. 대구: 태일사.

이경호, 고영만. 2002. 정보학, 서울: 인쇄마당.

임지룡. 1997. 인지의미론. 서울: 탑출판사.

정필모. 2004. 문헌분류론. 파주: 한국학술정보.

철학대사전. 1963. 서울: 학원사.

최석두. 2000. 시소러스 개발 지침. 서울: 한국데이터베이스진흥센터.

한유석, 설근수. 2004. 한국어 시소러스 연구. 서울: 한국문화사.

Dahlberg, I. 1978. "A Referent-Oriented, Analytical Cencept Theory for INTERCEPT". International Classification, 5(3): 142-151.

Garshol, Lars Marius. 2004. "Metadata? Thesauri? Taxonomies? Topic Maps!", [인용일자 2015. 7. 16] <http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#sect-taxonomies>

Koch, Traougott. 1997. "The Role of Classification Schemes in Internet Resource Description and Discovery". [인용일자 2015. 7.20] <http://www.ukoln.ac.uk/metadata/desire/classification/class_1.htm>

Kwasnik, Barbara H. 1999. "The Role of Classification in Knowledge Representation and Discovery", Library Trends, 48(1): 22-47

ISO/IEC 2005. International Standard ISO/IEC 11179-2. Information technology -Metadata registries(MDR)- Part 2: Classification. 2nd ed. Geneva: ISO.

Noy, Natalya F. and Deborah L. McGuinness. 2008. Ontology Development 101: A Guide to Creating Your First Ontology. [인용일자 2015. 7. 17] <http://protege_stanford.edu/publications/ontology_development/ontology

101-noy-mcguinness.html>

Oxford English Dictionary: The definitive record of the English language. [인용일자 2015. 7. 16] <http://www.oed.com.ca.skku.edu:8080/>

Pidcock, Woody. 2003 "What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model? [인용일자 2015. 7. 16] <http://infogrid.org/trac/wiki/Reference/PidcockArticle>

Schmitz-Esser. Winfried. 1991. "New Approaches in Thesaurus Application". International Classification, 18(3): 143-147.

Taylor, John R. 1997. 인지언어학이란 무엇인가?: 언어학과 원형이론. 조명원, 나익주 옮김. 서울: 한국문화사.

The UDC: Essays for new decade. edited by Alan Gilchrist and David Strachan. 1990. London: Aslib.

Wikipedia. [인용일자 2015. 7. 16] <https://en.wikipedia.org/wiki/Classification>

Chan, Lois Mai. 1999. A Guide to the Library of Congress Classification. 5th ed. Englewood: Libraries Unlimited.

Hwketu, Meron. 2011. "The UNESCO Thesaurus". Paris: UN-LINKS Meeting.

Hunter, Eric J. 2015. 분류란 무엇인가: 지식의 구조화와 검색에 관한 이해. 박지영 옮김. 파주: 한울.

Iglesias-Sucasas, Marta, Soonho Kim and Virginie Viollier. "The FAO Geopolitical Ontology: reference for country-based information". [인용일자 2015. 8.20] <http://www.google.co.kr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&ved=0CDAQFjACahUKEwjP8M655MrHAhUiHqYKHdOiBEY&url=http%3A%2F%2Fwww.semantic-web-journal.net%2Fsites%2Fdefault%2Ffiles%2Fswj179_0.pdf&ei=qs7fVc_zCKK8mAXTxZKwBA&usg=AFQjCNG2PLwT6PrT7Uqh_L3mTd_9qgQmpQ&cad=rjt>

<저자 소개>

서태설 KISTI 정보서비스실 책임연구원
tsseo@kisti.re.kr
김비연 성균관대학교 문헌정보학과 교수
korkby@gmail.com

서비스를 위한 분류, 시소러스, 온톨로지의 이론과 사례

2015년 10월 29일 인쇄
2015년 10월 29일 발행

발 행 처

대전광역시 유성구 대학로 245
㉾ 305-806
전화 : 042-869-1004
등록 : 1991년 2월 12일 제 5-259호

발행인

인쇄처