# 시소러스 기반 온톨로지에 관한 연구*

## A Study on the Ontology based on Thesaurus

고 영 만(Young-Man Ko)**

목 차



### 초 록

본 연구에서는 시소러스의 의미 체계가 지니고 있는 한계 및 이를 극복하기 위한 방안들을 분석하였다. 분석은 최근 활발하게 논의되고 있는 온톨로지 개념 도입에 의한 시소러스의 개선 연구에 초점을 맞추었으며, 아울러 이러한 방안에 내재된 문제점과 한계 및 발전 방향을 전망하였다. 분석 결과 시소러스를 온톨로지로 발전시키는 작업의 성공은 최소한 해당 도메인에서 나타나는 개념 관계를 분석하여 일정한 패턴을 모두 찾아내고, 찾아낸 관계 유형과 그 유형을 표현하기 위한 복잡하고 수많은 술어를 어떻게 정형화시키고 표준화시키는가에 달려있는 것으로 나타났다. 그리고 온톨로지를 구축하기 위한 온톨로지 언어의 완성도 향상과 표준화 및 이를 토대로 하는 의미 관계 유형과 술어의 표준화 작업은 상당한 기간이 소요될 것으로 전망되었다.

### ABSTRACT

The week point of the thesaurus is characterized by the ambiguity and inconsistency of the semantic relationships. In this study, several studies which tried to improve the thesaurus applying the ontology principles were analysed and the problems and limitations of the methods were prospected. The result of analysis shows that the success of constructing ontology from thesaurus depends on the finding out of every patterns and predicates of semantic relationships in a domain specific thesaurus and the standardization of the patterns and predicates.

키워드: 시소러스, 시소러스 패싯, 온톨로지, 온톨로지 템플릿, 온톨로지 에디터, 의미검색, 추론
Thesaurus, Thesaurus facet, Ontology, Ontology template, Ontology editor, Semantic retrieval, Inference

---

* 이 연구는 2005년도 국회도서관 시소러스 DB 유지관리사업의 일환으로 수행되었음.
** 성균관대학교 문과대학 문헌정보학과 교수(ymko@skku.edu)

# 1. 서 론

시소러스는 특정 주제 영역에서 사용되는 용어와 이들 용어간의 의미관계를 체계적으로 제시한 색인어휘집으로서 색인과 검색 과정에서 디스크립터와 검색어를 선정하기 위한 도구로 사용된다. 시소러스는 용어간의 의미관계를 제시하고 있다는 점에서 검색 효율성을 개선하는 기능을 가지고 있으며, 이에 따라 정보검색 보조도구로서의 시소러스 개발을 위한 많은 연구와 이를 실현하기 위한 노력들이 있었다.

검색 보조도구로서의 시소러스가 갖는 중요성 및 유용성이 강조되는 것에 비례하여 시소러스의 의미체계가 지니는 관계 구조의 단순성과 표현의 한계성에 대한 문제가 심각하게 제기되었으며 이를 개선하기 위한 노력들이 지속되어 왔다. 또한 여러 개의 데이터베이스 또는 여러 분야의 데이터베이스를 동시에 다루는 일이 빈번해지고, 특히 웹 분야에서의 의미적 검색과 호환성을 목표로 하는 시멘틱 웹에 대한 연구가 활발해짐에 따라 시소러스의 호환성 역시 중요한 연구 과제가 되었다.

최근에 발표되고 있는 시소러스의 개선을 위한 연구는 크게 두 갈래로 나누어진다. 하나는 시소러스가 가지고 있는 용어 간 의미관계 구조화의 제한성을 극복하고 관계 유형을 확장함으로써 호환성을 확보하거나 보조 온톨로지와 결합시킴으로써 추론을 가능하게 하는 방안이며, 다른 하나는 용어정의 시소러스를 구축함으로써 의미의 불명확성을 제거하여 호환성 문제를 극복하는 방안이다.

본 연구에서는 시소러스의 의미 체계가 지니고 있는 한계 및 이를 극복하기 위한 방안들을 분석함으로써 시소러스의 발전 방향과 그 전망을 밝히고자 한다. 이를 위해 최근 연구되고 있는 연구 사례를 분석하고, 이러한 시도들에 내재된 문제점과 한계를 전망하였다. 특히 본 연구에서는 정보기술의 발전과 정보검색 환경의 변화에 따라 최근 매우 활발하게 논의되고 있는 온톨로지 개념 도입에 의한 시소러스의 개선 방안에 초점을 맞추어 분석하였다.

# 2. 시소러스의 개념 간 관계 구조의 문제점

시소러스 용어의 의미 체계에서 가장 중요한 관계는 광의어(Broad Term : BT), 협의어(Narrow Term : NT), 관련어(Related Term : RT)로 표현되는 개념 간의 관계이다. 개념 간의 관계를 보조하는 장치로서 동의어나 유사동의어 또는 드문 경우이긴 하지만 반의어를 표현하기 위한 우선어/비우선어 관계를 사용하며, 용어의 다의성 문제를 해결하기 위해서는 한정어를 사용한다. 또한 용어의 분류를 위해 계층분류 방식이나 패싯을 사용하기도 한다(이재윤, 김태수 1998).

BT와 NT로 표현되는 계층관계는 집합과 요소와의 관계를 나타내는 속-종(generic) 관계, 신체조직이나 지리상의 위치 또는 학문 분야 등을 나타내는 부분-전체(part-whole) 관계, 카테고리와 그에 포함된 예시를 나타내는 사례(instance) 관계로 엄격하게 제한되는 것이 일반적이다. 그러나 RT로

표현되는 연관관계는 물체와 특성, 인과관계, 조작과 행위자, 개념과 측정 단위 등 계층적이지도 않고 등가 관계도 아니면서 상당한 관련성이 있는 관계를 모두 포함한다. 따라서 연관관계는 관계의 규정 범위가 지나치게 단순하고 포괄적이다. 시소러스를 개선하고자 하는 연구와 개발 시도의 대부분은 기본적으로 시소러스의 단순한 구조와 체계에 내재되어 있는 용어 간 의미 구조의 불명확성, 개념에 대한 본질적 속성 파악의 어려움, 의미론적 정보검색 및 추론 기능 제공의 어려움을 극복하기 위한 것이다.

### 2. 1 용어간 의미 구조의 불명확성

시소러스에서는 일반적으로 색인과 검색에서 사용하는 개념에 상응하는 우선어가 선택되고, 이 우선어에 상응하는 디스크립터가 선정된다. 디스크립터는 비디스크립터와 동의 관계에서만 연결될 수 있으며, 비디스크립터 간에는 어떠한 관계나 연결도 설정되지 않는 의미 구조상의 한계를 가진다. 예를 들면 키워드 방식의 검색엔진이 동형이의어를 구분하는 못하는 것처럼 시소러스를 구성하는 용어들도 계층구조는 가지고 있으나 광범위하고 기초적인 관계들로만 구조화되어 있어서 동형이의어나 특정 용어가 복수의 개념을 갖는 동의어 등을 구분하지 못하는 문제를 가지고 있다(유영준 2005, 126).

시소러스는 동의어, 유사동의어, 반의어의 관계 구조에서도 볼 수 있듯이 개념에 관한 정보와 용어에 관한 정보를 구분하지 않고 혼합해서 표현한다. 개념은 용어로 표현됨으로써 그 의미 지평과 용어가 연결된다는 점에서 개념 정보와 용어 정보를 구분하지 않는 시소러스 용어 구조에서는 그 의미 관계와 계층관계가 분명하지 못하다. 따라서 시소러스는 검색 대상이 되는 주제 영역 내에서 개념을 분명하게 식별해내지 못하는 문제점을 가지고 있다.

### 2. 2 개념에 대한 본질적 속성 파악의 어려움

색인도구로서의 시소러스가 검색 효율을 개선시키기 위해서는 개념의 표현 용어를 일관성 있고 정확하게 사용해야 하며 모든 용어 간의 의미 관계를 분명하게 제시할 수 있어야 한다. 용어간의 의미 관계를 명확하게 제시하기 위해서는 각 용어가 지닌 의미를 명확하게 하는 것이 전제되어야 한다. 대부분의 시소러스는 용어 간의 의미 관계를 명확하게 하기 위한 장치로서 용어가 지닌 의미 속성을 직접 제시하는 대신 용어 간의 관계구조나 범위주기(scope note)를 통한 간접적인 간접적 방식을 택하고 있다. 그러나 시소러스에 수록되는 용어는 그 범위를 한정할 수 없으나 용어의 양은 증가할 수밖에 없다(김태수 2001, 232). 따라서 시소러스의 단순한 관계 구조와 범위주기를 통해서는 시소러스에 수록된 개념의 본질적 속성 평가가 어렵게 되며, 개념 간의 충돌 증가와 관계 구조의 설정이 점점 어렵게 되는 결과로 이어진다.

### 2. 3 의미 검색 및 추론 기능 제공의 어려움

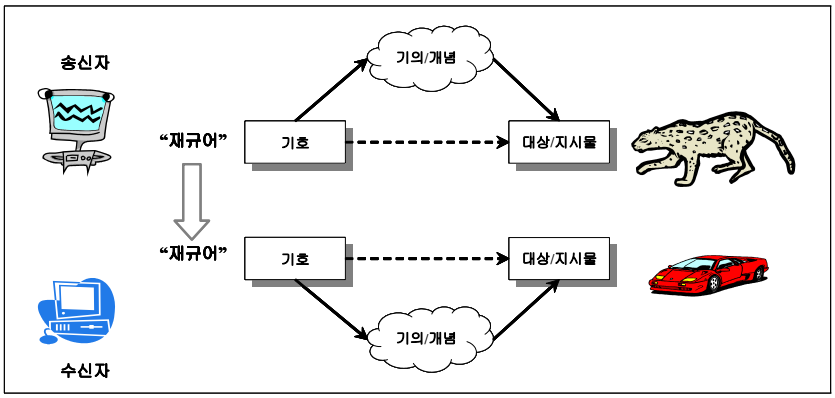
시소러스 용어의 개념 사이에서 나타나는 관계의 모호성은 시소러스의 단순한 관계 유형만으로는 개념간의 관계를 일관성 있게 유지시킬 수 없는 것에서 비롯된 것이다. 예를 들어 '국회도서관'이 '국회'의 하위어(NT)로 표현되는 시소러스의 관계 구조에서는 두 개념 간의 관계가 어떤 측면에서의 계층 관계인지 명시적으로 드러나지 않는다. 시소러스에서는 또한 계층 관계가 연관 관계로 구조화되거나 연관 관계가 계층 관계로 구조화되는 경우가 빈번하게 발생하며 이는 개념 간의 관계를 모호하게 만든다. 이러한 시소러스의 구조 속에서는 디스크립터 'A'와 디스크립터 'B' 각각의 디스크립터에 속해 있는 관련어 간에 상호 어떠한 연관성이 있는지를 알아내는 것이 거의 불가능하다.

개념을 계층 관계(BT/NT)와 연관 관계(RT)로 구분하여 개념 간의 관계를 구조화하고 있는 시소러스의 단순한 구조에서는 이용자들의 질의를 확장하거나 구체적으로 표현하기 위해 구축된 지식조직시스템의 개념 관계 구조화에 적용하는 것이 어려우며, 의미론적 정보 검색이나 추론 기능의 지원 역시 거의 불가능하다. 특히 웹 자원을 검색하거나 의미적으로 기술하는데 있어서는 시소러스의 관계 유형만으로는 적절하지 않다. 따라서 보다 다양하고 심층적인 개념 간의 관계 구조화를 토대로 하여 온톨로지 방법론을 적용하고자 하는 연구와 토론이 활발하게 이루어지고 있다.

## 3. 시소러스와 온톨로지

### 3. 1 온톨로지

전통적으로 온톨로지란 존재와 존재자(存在者 : 존재하는 것)의 본성을 연구하는 형이상학의 한 부분으로서 세상의 구성 요소에 대한 명확한 이해를 얻고자 하는 철학의 연구 분야이다. 오늘날 정보자원 관리와 관련해서 차용하고 있는 온톨로지(ontologly)라는 용어는 '사람의 마음 속에 존재하는 내재적 생각이나 외재적 세계의 현상과 대상에 대하여 공유하는 개념을 컴퓨터가 이해할 수 있는 형식으로 명확하고 명시적으로 정의하고 규정하는 것'으로 그 의미가 전이되어 사용되고 있다(그림 1 참조).




〈그림 1〉 대상에 대한 의미의 공유 문제

오늘날 정보자원 관리와 온톨로지의 연구는 대략 전문용어학(Terminology) 분야에서 관심을 갖는 전문분야 온톨로지, 데이터의 의미관리를 위한 데이터 및 메타데이터 온톨로지, 시맨틱웹 분야의 웹 온톨로지, 인공지능(AI) 분야의 의미망 온톨로지(Semantic Net Ontology) 등 네 가지 갈래로 나뉘어 수행되고 있다.(문헌)정보학의 연구 영역과 밀접하게 연관되는 전문용어학의 연구 분야인 시소러스 관점에서 온톨로지를 바라볼 경우 "보다 풍부한 지식(Knowledge-rich)을 갖추도록 용어의 의미 관계와 연결 정보를 보다 유동적이고 상세하게 기술화함으로써 의미추론 가능성과 시스템 간의 상호 운영성을 향상시키는 시소러스 관계 구조의 확장 개념"으로 이해되기도 한다.

### 3. 2 시소러스와 온톨로지의 차이

지식은 개념이 구조화된 것으로서 지식의 구조는 개념 간의 관계에 의해 규정된다. 용어 및 용어의 의미를 체계화시켜 지식을 구조화하는 도구로는 분류, 텍사노미, 시소러스, 온톨로지를 들 수 있다(표 1 참조). 특히 시소러스와 온톨로지는 개념 간의 의미 관계를 기반으로 지식의 구조를 정형적으로 표현하고 구조화한 것으로서 궁극적으로 검색 효율을 높이기 위한 측면에서 그 방법론과 구조에 대한 개발이 이루어졌다.

〈표 1〉 지식의 구조화 도구

| | | | | | |
|---|---|---|---|---|---|
| 통제용어<br>통제용어<br>통제용어 | +<br>+<br>+ | 그룹화<br>계층구조<br>용어관계 | □<br>□<br>□ | 분류<br>텍사노미<br>시소러스 | 용어<br>검색 |
| 통제용어<br>통제용어 | +<br>+ | 의미관계, 제한, 공리, 규칙<br>사례(instances) | □<br>□ | 온톨로지<br>지식베이스 | 의미<br>검색 |

온톨로지가 시소러스와 다른 점의 하나는 시소러스에 비해 개념 관계를 보다 세분하여 차별화할 수 있는 구조를 갖추고 있다는 점이다. 온톨로지는 개념 간의 관계와 용어 간의 관계를 분리하여 해당 주제 영역을 파악할 수 있는 구조를 갖추고 있다. 이를 통해 인간의 이해 구조를 더 잘 반영할 수 있으므로 시소러스에 비해 개념 간의 관계를 보다 정확하고 분명하게 만든다. 또한 주제 영역 내에서 일관성 있고 명확하게 각각의 개념을 정의하고, 개념 간의 관계를 구조화함으로써 해당 주제 영역의 특성을 보다 더 분명하게 반영할 수 있도록 해준다.

온톨로지가 시소러스와 구별되는 또 다른 특성은 일반화 혹은 상호운영성의 규칙을 적용함으로써 구조화된 지식으로부터 새로운 지식을 추론할 수 있다는 점이다. 추론을 통해 새롭게 덧붙여지는 지식은 지능적인 정보 처리에 적용될 경우 많은 역할을 할 수 있다.

표 2는 ERIC 시소러스의 관계 구조에서는 개별적인 연관관계로만 구조화되는 'literary education'과 'reading attitudes'가 온톨로지의 관계 구조에서는 상호 추론이 가능한 규칙 설정이 가능하게 되는 것을 보여준다(표 2 참조 : 유영준 2005, 128-129).

〈표 2〉 ERIC 시소러스와 온톨로지의 개념 관계 비교 및 온톨로지 의미관계에 의한 추론 규칙

| ERIC 시소러스 | 온톨로지 |
|---|---|
| reading instruction<br>BT instruction<br>RT reading<br>RT literary education | reading instruction<br>〈isa〉 instruction<br>〈hasDomain〉 reading<br>〈governedBy〉 literary education |
| reading ability<br>BT verbal ability<br>RT reading<br>RT reading attitudes | reading ability<br>〈isa〉 verbal ability<br>〈hasDomain〉 reading<br>〈supportedBy〉 reading attitudes |

| 규칙 1 |
|---|
| Instruction in a domain should consider ability in that domain:<br>X shouldConsider Y<br>IF X 〈isa(type of)〉 instruction AND X 〈hasDomain〉 W<br>AND Y 〈isa〉 ability AND Y 〈hasDomain〉 W<br>yields: The designer of reading instruction should also consider literary education. |
| 규칙 2 |
| X shouldConsider Z<br>IF X 〈shouldConsider〉 Y<br>AND Y 〈supportedBy〉 Z<br>yields: The designer of reading instruction should also consider reading attitudes. |

또 다른 예로서 표 3은 AGROVOC 시소러스의 관계 구조에서는 상호 관계를 맺지 못하는 'milk fat'와 'cheddar cheese'가 온톨로지의 관계 구조에서는 상호 추론이 가능한 관계가 될 수 있음을

보여 준다(표 3 참조 : 유영준 2005, 128-129).

〈표 3〉 AGROVOC 시소러스와 온톨로지의 개념간 관계 비교 및 온톨로지에 의한 추론 규칙

| AGROVOC | 온톨로지 |
|---|---|
| milk<br>NT cow milk<br>NT milk fat | milk<br>〈includesSpecific〉 cow milk<br>〈containsSubstance〉 milk fat |
| cow<br>NT cow milk | cow<br>〈hasComponent〉 cow milk |
| Cheddar cheese<br>BT cow milk | Cheddar cheese<br>〈madeFrom〉 cow milk |

| 규칙 1 |
|---|
| Part X 〈mayContainSubstance〉 Substance Y<br>IF Animal W 〈hasComponent〉 Part X<br>AND Animal W 〈ingests〉 Substance Y |
| 규칙 2 |
| Food Z 〈containsSubstance〉 Substance Y<br>IF Food Z 〈madeFrom〉 Part X<br>AND Part X 〈containsSubstance〉 Substance Y |

# 4. 온톨로지의 관계 구조 적용에 의한 시소러스 개선

온톨로지 개념을 도입하여 시소러스를 개선하고자 하는 연구는 크게 두 가지로 나누어 볼 수 있다. 하나는 기존 시소러스의 패싯과 온톨로지의 개념 템플릿을 연결하는 연구이며 다른 하나는 시소러스의 개념 간 관계를 상세하게 분석하여 확장하거나 온톨로지 개념 구조와의 상호호환 가능성을 모색하는 방식에 관한 연구이다.

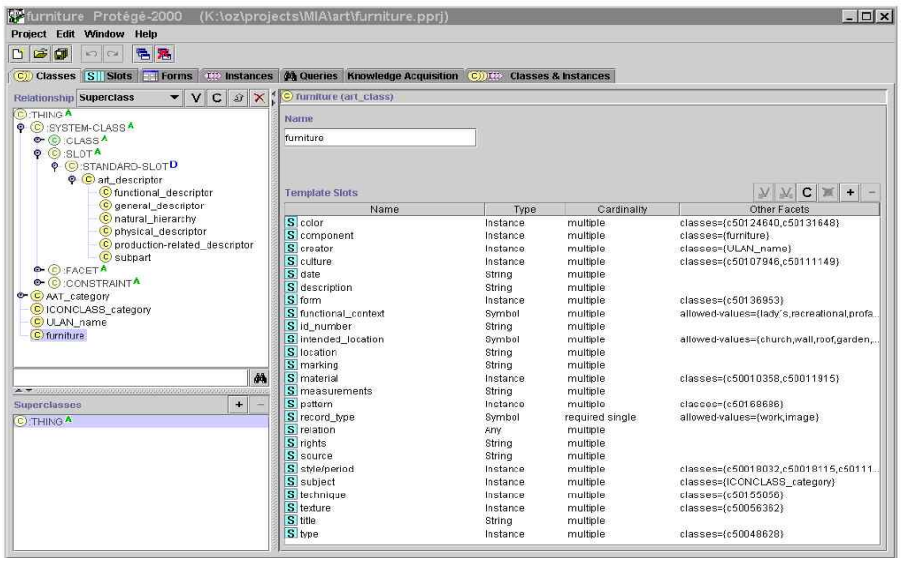
## 4. 1 시소러스 패싯과 온톨로지 템플릿 연결

이 방식은 온톨로지의 개념 표현 템플릿과 기존 시소러스의 패싯을 연결하여 시소러스에 포함된 지식 구조를 보조 온톨로지로 사용함으로써 추론 가능성을 높이는 것이다. 어떤 개념을 표현하는데 있어서 풍부한 지식을 부여함으로써 개념 간의 추론이 가능한 시스템을 구현하기 위하여 기존 시소러스의 패싯과 어휘 구조를 이용한다. 개념을 표현하기 위해 구성하는 온톨로지 템플릿은 기본적으로 해당 분야 분류체계의 대분류 항목 또는 시소러스의 패싯을 사용하며, 추가적으로 다른 목적으

로 개발된 메타데이터 요소들을 사용하기도 한다. 템플릿 각각의 요소는 시소러스 패싯과의 연결을 통해 그 시소러스의 어휘구조를 이용하게 되며, 이때 추출된 개념의 계층 구조에 WordNet이나 다른 문헌과 같은 다양한 자원으로부터 가져온 관련 지식을 추가함으로써 보다 풍부한 지식을 가지는 개념이 생성된다.

시소러스 패싯과 보조적인 요소들로 지식베이스의 템플릿을 구성하여 지식베이스 시스템을 구축한 대표적인 연구 사례로는 빌링가(Wielinga 2001) 등의 연구를 들 수 있다. 이들은 골동가구 분야의 지식베이스를 구축하기 위하여 VRA(Visual Resource Association)의 핵심범주(Core Category)와 더블린코어 메타데이터 요소 및 European GRASP 프로젝트 결과로부터 추출한 25개 요소로 템플릿을 구축하였으며(그림 2 참조), 개개의 골동가구 단위는 구조화되지 25개의 요소(속성)로 기술되며, 25개의 속성들은 다음과 같은 4개 패싯으로 나뉘어 구조화하였다 :

- 제작관련 패싯 : creater, style/period, technique, etc..
- 물리적 패싯: measurements, color, material, etc.
- 기능적 패싯: intended location functional context
- 관리 패싯 : collection ID, rights, current location.




〈그림 2〉 골동가구 기술 템플릿

템플릿의 구조화된 패싯을 해당 분야의 대표적 시소러스인 AAT(Art and Architecture Thesaurus)의 패싯과 상호 연결시킴으로써 AAT의 어휘구조를 활용할 수 있는 기반을 구축하였으며, 온톨로지 에디터 Protege-2000을 사용하여 AAT의 계층관계를 온톨로지 구조에 활용하였다.

## 4. 2 시소러스의 개념 간 관계 확장

유영준(2005)은 "온톨로지 개념간 관계 설정을 위한 AGROVOC 시소러스의 분석에 관한 연구"에서 온톨로지 개념을 도입하여 시소러스 개념 간의 관계를 확장함으로써 시소러스의 한계를 극복하는 시도를 하였다. 이 연구는 시소러스를 웹상에서 효율적으로 활용하기 위해서는 개념 간의 관계를 보다 정밀하게 정의하고, 해당 개념을 표현하는 용어를 구조화한 주제 영역의 개념들로 시소러스가 조직되어야 하며, 이와 함께 연관된 개념들을 일반화하고 추론 기능이 가능하도록 타당성 있는 규칙과 조건들이 구체적으로 제시되어야 함을 전제로 한 것이다. 유영준의 연구에서는 시소러스의 개념 간 관계 분석을 통해 온톨로지의 핵심인 의미론적으로 보다 발전된 개념 간 관계 유형을 제시하고 있다.

### 4. 1. 1 개념, 용어, 문자열의 분리 모형

유영준이 UMLS의 개념 구조를 기반으로 시소러스의 의미관계를 분석하여 제시한 모형에서의 개념(concept)은 용어(term)나 어휘(lexicalisation)에 의해 표현되거나 지시되며 이들은 단일어이거나 구 또는 복합어일 수 있다. 개념은 계층에 따라 배열되고 네트워크 내의 다른 개념들과 관계를 맺으며, 특정한 경우에는 〈componentOf〉와 〈hasComponent〉 관계와 같은 역관계를 가질 수 있다. 온톨로지에서의 개념간 관계유형을 결정하는 기반은 관련 시소러스에 존재하는 개념간 관계를 상세하게 분석해서 도출된 관계 유형이다. 따라서 시소러스의 개념간 관계 구조에 대한 지식이 매우 중요하며, 이러한 구조 속에서의 각 개념은 하나의 엔터티 유형 또는 패싯으로 간주될 수 있다.

개념은 용어(term/lexicalisation)로 표현됨으로써 그 개념의 수준과 용어의 수준이 연결된다. 용어는 단수와 복수, 격변화, 철자변형, 약어, 두문자어와 간은 다양한 어형 변화를 가질 구 있다. 따라서 하나의 개념은 다양한 어휘 표현을 가질 수 있으며 용어는 다양한 기호를 가질 수 있다. 이때 동형이의어나 동의어, 또는 특정 용어가 복수의 개념을 갖는 경우에는 혼란을 피하기 위하여 개념들이 독립적으로 존재해야 하며, 각 용어의 의미 속성이 분명한 용어들만 상응하는 개념과 연결되어야 한다. 따라서 개념 간 관계를 구조화하는 시소러스나 온톨로지에서는 개념과 용어를 분리시켜서 모형화하는 것이 타당하다고 할 수 있다.

개념에 대한 정보와 용어에 대한 정보를 분리시킬 경우 시소러스에서는 이용자가 선호하는 개념에 상응하는 디스크립터의 선택을 탄력적으로 적용할 수 있으며, 〈hasSynonym〉, 〈hasAntonym〉, 〈hasCognate〉, 〈hasTranslation〉과 같은 다양한 관계로 용어를 연결시킬 수 있게 된다. 용어(term)는 문자열(String)로 표현되며, 문자열은 〈hasCaseVariant〉, 〈hasSpellingVariant〉, 〈pluralOf〉, 〈singularOf〉, 〈hasAbbreviationOrAcronym〉와 같은 관계를 통해서 연결될 수 있게 된다. 따라서 시소러스는 그 용어를 표현하기 위해 사용된 문자열 중에서 선호하는 이형〈variant〉을 선택하면 되고, 특히 반의어에서 하나의 문자열은 여러 용어들에 동시에 포함될 수 있다. 이때 개념, 용어, 문자열을 독립된 엔티티나 패싯유형으로 정의할 경우 엔티티/패싯 유형 각각은 상이한 유형의 정보를 가질 수 있으며

이를 통해 혼란을 피할 수 있게 된다.

4. 1. 2 시소러스 개념 관계의 세분화

유영준은 AGROVOC 시소러스의 개념 간 관계를 분석함으로써 온톨로지에 적용할 수 있도록 보다 세밀하고 분명하게 기술할 수 있는 방안을 제시하고 있다. 제시된 관계 유형 역시 기존의 시소러스의 기본 관계인 종속관계, 부분-전체관계, 추가관계를 토대로 세분화한 것으로서 이를 통해 시소러스를 온톨로지 수준으로 발전시키고자 한 실험적 시도라 할 수 있다. 유영준의 연구에서 도출된 시소러스 개념 간 관계는 다음과 같다 :

가) 종속관계 : NT관계의 세분화

- X〈includesSpecific〉Y ↔ Y〈isa〉X
  X〈inheritsTo〉Y ↔ Y〈inheritsFrom〉X

나) 부분-전체관계 : NT 관계의 세분화

- X〈containsSubstance〉Y ↔ Y〈substanceContainedIn〉X
  X〈hasIngredient〉Y ↔ Y〈ingredientOf〉X
- X〈yieldsPortion〉Y ↔ Y〈portionOf〉X
- X〈spatiallyIncludes〉Y ↔ Y〈spatiallyIncludedIn〉X
- X〈hasComponent〉Y ↔ Y〈componentOf〉X
- X〈includesSubprocess〉Y ↔ Y〈subprocessOf〉X
- X〈hasMember〉Y ↔ Y〈memberOf〉X

다) 추가관계

- X〈causes〉Y ↔ Y〈causedBy〉X
- X〈instrumentFor〉Y ↔ Y〈performedByInstrument〉X
- X〈processFor〉Y ↔ Y〈usesProcess〉X
- X〈beneficalFor〉Y ↔ Y〈beneficialFrom〉X
- X〈treatmentFor〉Y ↔ Y〈treatedWith〉X
- X〈harmfulFor〉Y ↔ Y〈harmedBy〉X
- X〈growsIn〉Y ↔ Y〈growthEnvironmentFor〉X
- X〈hasProperty〉Y ↔ Y〈propertyOf〉X
- X〈similarTo〉Y ↔ Y〈similarTo〉X
- X〈oppositeTo〉Y ↔ Y〈oppositeTo〉X

### 4. 2 시소러스와 온톨로지의 상호호환성 추구

이 방식은 정보검색에 있어서의 검색 효율성을 향상시키기 위하여 기존 시소러스의 독립성을 유지하면서 논리적 추론이나 통합검색이 가능하도록 온톨로지 개념을 적용하는 것이다. 두 개의 상이한 데이터베이스에서 같은 개념에 대해 사용하는 용어가 다를 경우 이 용어들이 동의어 관계에 있다는 것을 확인해 주는 방식으로 용어간의 관계를 이용하여 복수의 데이터베이스를 검색할 수 있는 알고리즘을 개발하는 것이 목적이다. 이러한 시도는 시소러스의 기반으로 사용하는 용어간의 개념 관계 구조화 방식과 온톨로지에서 사용하는 개념 표현 용어의 구조화 모형을 비교하여 유사성을 찾아내고 이를 통해 상호호환이 가능한 시소러스를 구축하는 방식으로 이루어진다.

조현양과 남영준(2004)은 '시소러스와 온톨로지의 상호호환성에 관한 연구'에서 시소러스의 개념 관계 구조와 온톨로지 구축 언어의 하나인 OWL Lite의 개념 관계 구조를 비교하여 분석하였다. 이 분석에서는 계층 관계의 경우 〈subClassOf〉와 〈subPropertyOf〉 등이, 연관 관계에서는 〈ObjectProperty〉, 〈DatatypeProperty〉, 〈inverseOf〉 등이, 대등 관계에서는 〈equivalent Class〉를 비롯하여 〈sameAs〉와 〈equivalentProperty〉 등이 적합한 것으로 나타났다.

가) 계층관계

- 속-종관계 : 〈Class〉 ↔ 〈subClassOf〉
- 전체-부분관계 : 〈Class〉 ↔ 〈subClassOf〉
- 사례관계 : 〈Class〉 ↔ 〈subClassOf〉
- 다계층관계 : 〈Class〉 ↔ 〈subClassOf〉
- 계층내 노드 레이블 : 〈ObjectProperty〉 ↔ 〈subPropertyOf〉

나) 연관관계

- 동일 범주내 중복 자매어 : 〈equivalentClass〉
- 동일 범주내 상호 독점적 자매어 : 〈differentFrom〉
- 동일 범주내 파생 관계 : 〈ObjectProperty〉
- 다른 범주의 용어간 관계 : 〈ObjectProperty〉
- 관련어용 노드 레이블 : 〈ObjectProperty〉〈DatatypePropertyOf〉

다) 대등관계

- 일반적 대등관계 : 〈sameAs〉
- 동의어 관계 : 〈sameAs〉
- 변형어휘 관계 : 〈sameAs〉

- 유사 동의어 관계 : 〈sameAs〉
- 복합명사에서 상호 참조 관계 : 〈equivalentProperty〉

# 5. 용어정의에 의한 시소러스 개선

## 5. 1 용어정의 시소러스

시소러스에 디스크립터의 정의를 도입하기 위한 구상은 무어스(Moors 1963), 쇠르겔(Soergel 1974) 등에 의해 제안되었으며, 용어학 분야에서는 전문용어와 시소러스를 연결한 용어 시소러스를 개발하는 시도를 하였다. 특정 주제 영역에 대한 정의 모형 적용 연구(Strehlow 1983)가 발표된 이후 시소러스 용어의 정의를 위한 정의 모형들이 사거와 롬(Sager and L'Homme 1994)와 허든(Hudon 1996) 등에 의해 제안되었다. 정의모형을 적용하여 사회과학 분야의 시소러스에 적용한 연구 성과가 1996년 사거와 롬에 의해 발표되었으며 1999년에는 정의 모형을 적용한 'WordWeb 시소러스(http://www.wordweb.co.uk)' 등이 개발되었다.

국내에서는 김태수(2001)에 의해 시소러스에 용어 정의를 이용하는 연구가 이루어진 바 있다. 김태수는 "용어정의를 도입한 시소러스 개발 연구"에서 사거와 롬의 정의 모형과 이 모형을 확장한 허든의 모형을 적용하여 정의 모형과 적용지침을 제시하였다. 제시한 정의 모형의 구성 요소 중에서 시소러스 구축에 필요하다고 판단되는 요소를 추출하여 이를 디스크립터 간의 관계 구조에 반영한 시소러스를 실험적으로 구축하였으며, 이 연구를 통해 의미범위와 관계구조의 표준화 가능성이 있음을 검증하였다.

사거와 롬이 제안한 용어 정의 모형은 전통적인 분석적 정의를 표준 형식으로 표현하기 위해 정의 기술 방식을 범주화한 것으로 주제분야, 피정의항의 개념범주, 정의항, 정의항의 개념범주, 피정의항과 정의항의 관계, 피정의항의 본질적 구별 특성(종차), 기타 특성의 일곱 가지의 요소로 구성되어 있으며, 김태수가 시소러스 구축을 위해 추출한 정의 모형 요소는 주제분야, 정의항, 피정의항과 정의항내 용어와의 관계, 구분특성(종차)에 의한 패싯 적용의 네 가지이다(표 4 참조).

〈표 4〉 Sager and L'Homme와 김태수의 정의모형 요소

| Sager and L'Homme의 정의모형 | | 김태수의 정의모형 | |
|---|---|---|---|
| 주제분야 | | 주제분야 | |
| 피정의항 개념범주 | · 물질(material)<br>· 추상(abstract)<br>· 활동(activity)<br>· 상태(state)<br>· 성질(property) | | |
| 정의항 | · 일반적으로 피정의항과 동일한 개념범주에 해당<br>· 피정의항과 가장 밀접한 개념<br>· 특수한 정의 개념의 경우 정의항 자체에 대한 간략한 정의 및 정의 확장 가능<br>· 정의하는 개념을 두 가지 이상으로 지시하는 것 회피<br>· 하나의 개념으로 정의가 불가능한 경우 복수의 정의한 사용 | 정의항 | · 상위 · 하위 개념<br>· 한정된 상위 개념<br>· 관련 단어나 구를 사용한 동일 수준의 추상적 개념 |
| 정의항의 개념 범주 | | | |
| 피정의항과 정의항의 관계 | · 피정의항이 정의항에 제시된 상위개념의 유형, 범주, 사례<br>· 피정의항이 하위개념<br>· 피정의항이 상위개념<br>· 피정의항이 정의항과 추상화 수준이 동일 | 피정의항과 정의항 내 용어와의 관계 | · 피정의항이 하위 개념<br>· 피정의항이 상위 개념<br>· 피정의항과 정의항의 용어가 동일 수준의 개념 |
| 피정의항의 본질적 구별 특성(종차) | · 본질적 특성(포함/구성, 소유/속성, 선행수식어/~의/~인/~된)<br>· 원인이나 기원 특성<br>· 변형이나 수정, 증가 등의 상태변화 특성<br>· 용도 특성<br>· 기능 또는 행위 특성<br>· 위치 특성<br>· 시간이나 연대 특성<br>· 유사성 특성 | 구분특성(종차)에 의한 패싯 적용 | · 본질적 특성<br>· 기원 특성<br>· 상태/변화 특성<br>· 목적/용도 특성<br>· 기능/행위 특성<br>· 위치 특성<br>· 시간 특성<br>· 유사 특성 |
| 기타 특성 | · 피정의항의 적용 범위나 정의항의 범위를 제한하거나 사례 제시 | | |

## 5. 2 용어정의 구조로서의 데이터 레지스트리

용어정의 시소러스는 시소러스 용어 하나하나에 대하여 구조적이고 상세한 정의를 한다는 점에서 넓은 의미의 데이터 레지스트리를 구축하는 것이라 할 수 있다. 시소러스 용어에 대한 데이터 레지스트리의 구조와 관리가 표준화될 경우 용어의 의미가 명확해질 뿐 아니라 레스트리 구조를 이루고 있는 다양한 요소 항목에 의한 검색과 추론이 가능하게 된다.

데이터 레지스트리(Data Registry)는 데이터를 등록하고 관리하기 위한 정보시스템이다. 데이터의 등록과 인증을 통하여 표준화된 데이터를 유지 관리하며, 데이터의 명세와 의미의 공유를 통하여

호환성을 높이는 것을 그 목적으로 한다. 데이터 레지스트리는 "공유되는 개념의 정형적 명시적 명세화 도구" 즉 온톨로지를 통해 데이터 간의 호환성을 유지시킨다. 구체적으로 표현하면 동일한 의미를 가지는 항목에 대해 서로 다른 표현(용어)을 사용할 경우 이를 해결해 주기 위해 공유되는 개념화를 정형적, 명시적으로 명세화한 결과물들이 집합되어 있는 데이터(혹은 용어)의 표준관리 시스템이라 할 수 있다. 예를 들어 국사와 한국사를 사용하는 두 개의 데이터베이스에 들어있는 데이터(또는 용어)를 비교하거나 통합하려는 프로그램에서는 국사와 한국사가 같은 의미를 지칭하는 데이터/용어라는 것을 알아야 하며, 데이터 레지스트리에는 바로 이러한 관계가 정형적, 명시적으로 명세화 된다.

국제표준화기구의 공동기술위원회(ISO/IEC JTC1)에서 데이터 교환 및 관리 표준화를 담당하는 분과위원회 SC32는 데이터의 의미, 구문, 표현을 표준화하기 위한 프레임워크를 개발하여 국제표준 "ISO/IEC 11179 - Metadata Registry"를 제정하였다. 이 표준은 본래 메타데이터의 등록과 인증을 통하여 표준화된 메타데이터를 유지・관리하게 함으로써 메타데이터의 명세와 의미를 공유할 수 있게 해주는 기능을 가진다(남영광 등 2005). 그렇지만 실제에 있어서는 데이터 레지스트리의 모형으로도 적용되고 있으며, 김태수의 정의 모형은 시소러스 데이터(용어) 레지스트리 모형으로 기능할 수 있는 가능성을 보여주는 것이라 할 수 있다.

## 6. 온톨로지 기반 시소러스의 발전 전망

시소러스의 개념 간 관계를 구조화하는 수단인 동의 관계, 계층 관계, 연관 관계의 유형만으로 의미 관계를 구조화하는 것은 지식의 연관성을 제대로 표현해내는데 있어서 완성도가 떨어진다. 온톨로지는 이러한 단점을 보완하기 위한 대안으로 고안된 것으로서 시소러스의 개념과 관계 유형을 분석해서 새로운 관계 유형을 찾아내고 기존의 단순한 관계 유형(BT, NT, RT 유형)을 보완함으로써 해당 분야의 온톨로지로 발전시키고 시소러스에서 실현하지 못했던 추론 규칙을 생성할 수 있게 하는 다양한 연구가 진행되고 있다.

시소러스의 개념 간 관계의 확장을 통해 온톨로지로 발전시키고자 하는 유영준의 연구와 온톨로지에서 사용하는 의미 관계를 시소러스에 적용하는 방식을 통해 시소러스의 한계와 문제점을 극복하고자 한 남영준, 조현양의 연구는 기본적으로 시소러스의 개념 관계 유형을 세분화한 다음 이를 온톨로지 기술 형식(언어, 도구)으로 표현하려고 한 점에서 방법론적 측면에서 거의 유사한 방향의 연구라 할 수 있다. 빌링가 등의 연구는 기본적으로 온톨로지 에디터를 사용하여 서술하는 해당 도메인의 용어에 관련 분야 시소러스를 연결하여 추론 가능성을 높이는 방안을 제시하고 있다. 이들은 국제적 표준으로 인정받고 있는 일반적 메타데이터 요소와 해당 도메인 분야의 메타데이터 요소로부터 패싯을 추출하고 이를 해당 도메인 분야의 기존 시소러스 패싯과 연결시킴으로써 용어에 보다 풍부한 지식을 부여하여 추론 가능성을 높이고자 하였다.

기존 시소러스 용어를 기반으로 그 관계 유형을 분석하여 온톨로지 서술 방식으로 발전시키고자 한 유영준, 남영준, 빌링가 등의 연구와는 달리 김태수의 연구는 기존 시소러스 용어 하나하나에 대한 명확하고 구조화된 정의를 통해 시소러스의 용어 간 관계의 모호성을 극복하고자 한 것이다. 시소러스 용어가 구조화된 정의 모형에 따라 명세화될 경우 모형을 구성하고 있는 요소(주제분야, 패싯, 특성 등)별 연관 관계를 추론하는 것이 가능하게 될 것이며, 이러한 명세화가 축적될 경우 궁극적으로는 매우 높은 수준의 온톨로지가 구축되는 결과가 될 것이다. 다만 시소러스가 용어정의 모형을 통해 온톨로지로 발전하기 위해서는 용어정의 모형의 표준화 및 용어정의 모형을 시스템적으로 관리할 수 있는 표준화 작업이 선행되어야 할 것이다.

온톨로지는 기본적으로 웹 자원의 의미 검색을 가능하게 하기 위한 수단의 하나로서 고안된 것이다. 온톨로지가 표준화된 웹자원의 조직 수단으로의 기능을 제대로 수행하기 위해서는 해당 도메인(주제 분야)의 용어들이 가지는 개념간의 관계를 분석해서 다양한 연관 관계 유형을 밝혀내는 작업이 선행되어야 한다. 따라서 시소러스를 온톨로지와 같은 구조로 변형시키기 위해서는 시소러스의 의미 관계들을 보다 명확하게 표현할 수 있도록 만들어야 하며, 이때 가장 중요한 것은 의미 관계에 어떠한 패턴이 있는지를 분석해서 일정한 패턴을 밝혀내는 일이다. 해당 도메인의 시소러스가 구축되어 있을 경우에는 시소러스가 포함하고 있는 용어들의 관계 분석을 통해 이를 온톨로지로 발전시키는 것이 비교적 용이할 것이며, 온톨로지를 새롭게 개발하는 것에 비해 훨씬 실용적이다. 그러나 해당 도메인 분야의 시소러스가 구축되어 있지 않을 경우 개념 간의 모든 관계 유형을 찾아내는 작업은 고도의 지적 노력이 요구되는 작업이 될 것이다.

관계 유형이 찾아지면 이를 정형화하는 작업이 필요하며, 정형적 관계 유형이 결정된 후에는 관계를 표현하기 위한 술어의 정형화 및 표준화 작업이 반드시 이루어져야 한다. 웹 자원의 의미 검색은 궁극적으로 모든 도메인의 관계 유형과 이 유형을 표현하는 술어가 표준화되어야 가능하기 때문이다. 상호운영성을 고려할 경우 관계 유형의 정형화는 최소한 해당 도메인 내에서 구축되는 모든 온톨로지에서는 동일하게 사용할 수 있어야 한다. 그렇지 않을 경우 온톨로지의 궁극적 목적인 의미 검색과 추론 규칙의 생성이 거의 불가능해 질 것이며, 매핑 작업과 같은 번거롭고 추가적인 보완 작업과 새로운 표준화 논의가 끊임없이 반복될 것이다.

시소러스를 온톨로지로 발전시키는 작업의 성공은 모든 도메인에서 나타나는 개념 관계를 분석하여 일정한 패턴을 모두 찾아내고, 찾아낸 관계 유형과 그 유형을 표현하기 위한 복잡하고 수많은 술어를 어떻게 정형화시키고 표준화시키는가에 달려있다고 할 수 있다. 그러나 온톨로지를 구축하기 위한 도구의 하나인 온톨로지 언어가 아직 표준화되어 있지 않아 모든 도메인에 해당되는 용어의 관계 유형과 술어를 표준화하는 작업은 상당한 기간이 소요될 것이다. 따라서 시소러스가 구축되어 있는 도메인과 이를 관리하는 기관에서는 우선 구축되어 있는 해당 시소러스에서 용어 관계의 오류를 바로잡아 일관성을 유지시키기 위한 작업을 수행하는 것이 필요하다.

# 참 고 문 헌

김태수. 2001. 용어 정의를 도입한 시소러스 개발 연구. 『정보관리학회지』, 18(2): 231-254.

남영광, 서태설, 황상원. 2005. ISO/IEC 11179 표준에 따른 산업기술정보 메타데이터 표준화. 『정보관리연구』, 36(1): 57-75.

유영준. 2005. 온톨로지의 개념간 관계 설정을 위한 AGROVOC 시소러스의 분석에 관한 연구. 『정보관리학회지』, 22(1): 125-144.

이재윤, 김태수. 1998. WordNet과 시소러스. 『제11회 언어정보연찬회 발표 논문집』, 1998.2.10, 연세대학교.

조현양, 남영준. 2004. 시소러스와 온톨로지의 상호 호환성에 관한 연구. 『정보관리학회지』, 21(4): 27-65.

Hudon, M. 1992. "Term Definitions in Subject Thesauri : The Canadian Literacy Thesaurus Experience." Classification Research for Knowledge Representation and Organization. Edited by Williamson, N. J. and M. Hudon. Amsterdam : Elsevier. 255-262.

Moors, C. N. 1963. "The Indexing Language of an Information Retriefal System." *Information Retrieval Today : Papers presented at an Institute conducted by the Library School and the Center for Continuation Study, University of Minnesota.* Ed. by W. Simonton. Minneapolis, MN : The Center. 21-36

Sager, J. C. and M.-C. L'Homme. 1994. "A Model for the Definition of Concepts : Rules for Analytical Definitions in Terminological Databases." *Terminology,* 1(2): 61-81.

Soergel, D. 1974. *Indexing Languages and Thesauri : Construction and Maintnance*. Los Angeles, CA : Melville.

Strehlow, R.A. 1983. "Terminology and the Well-formed Definition." *Standardization of Technical Terminology : Principles and Practices*, 806. 15-25. Edited by Interrante, C. G., and F. J. Heymann, Philadelphia, PA : American Society for Testing and Materials.

Wielinga, B.J, A. Th. Schreiber, J. Wielemaker, J. A. C. Sandberg. 2001. From Thesaurus to Ontology. Proceedings of the 1st International Conference on Knowledge Capture. Oct. 22-23, 2001, Victoria, BC, Canada. 194-201.