하이퍼텍스트를 이용한 온라인 시소러스의 선형배열 설계에 관한 연구

A Study on the Design of Hypertext-Based Linear Displays for an Online Thesaurus

최 재 황(Jae-Hwang Choi)*

목 차

초 록

본 연구의 목적은 ISO와 ANSI/NISO의 시소러스 작성지침을 참고하여 문헌정보학분야 시소러스를 하이퍼텍스트를 이용하여 선형배열로 웹 상에 설계해 보는데 있다. 본 연구는 하이퍼텍스트를 이용한 온라인 시소러스가 정보를 저장하고 탐색하는 사람들에게 편리하고 유용한 주제접근의 도구가 되며, 인쇄형 시소러스로는 파악하기 어려운 이용자의 시소러스 탐색유형을 연구하는 기초가 될 수 있을 것이라는 가정에서 출발하였다. 본 연구를 위해서 문헌정보학 분야 시소러스를 관계형 데이터베이스인 MS ACCESS 97에 저장하였고, 관계형 데이터베이스와 웹과의 연동을 위해서 Windows NT운영체재 하에서 ASP(Active Server Pages)기술을 적용하였다.

ABSTRACTS

The purpose of this study is to design hypertext-based linear displays for an online thesaurus in librarianship and information science with the aid of ISO and ANSI/NISO thesaurus standards. This study starts with the assumptions that hypertext-based online thesauri would provide a convenient and useful subject retrieval tool to both indexers and searchers of information and become starting point for the study of thesauri searching patterns, which were difficult with printed thesauri. For this study, thesaurus of librarianship and information science was stored in MS ACCESS 97 as a relational database and, for the conjunction of a relational database with World Wide Web, technics of ASP(Active Server Pages) were applied under Windows NT operation.

* 성균관대학교 문헌정보학과 강사
접수일자 1999년 8월 20일

1. 서 론

우리는 시소러스를 흔히 용어의 집합, 관계의 집합, 그리고 배열(표시)의 집합이라고 말한다. 용어의 집합에는 디스크립터(색인어로 선정된 단어 또는 구)와 비디스크립터(색인어로 선정되지 않은 단어 또는 구)가 포함되고, 관계의 집합에는 용어간의 관계인 동등관계, 계층관계, 연관관계를 포함하게 된다. 용어간의 관계는 하나 이상의 배열로 보여지게 되는데 여기에는 알파벳순 배열, 계층 배열, 체계 배열, 도식배열 등을 포함하게 된다. 알파벳순 배열, 계층 배열, 체계 배열 등은 선형배열(linear displays)이라 하여, 도식배열(graphic displays)과는 구별된다.

시소러스는 정해진 순서와 구조로 배열된 색인어의 통제어휘집을 말한다. 시소러스에 의해서 용어간의 동등관계, 동형이의어 관계, 계층관계, 연관관계가 분명해 지고, 이들은 표준화된 관계기호에 의해서 식별된다. 시소러스는 색인어간의 어의적 관계뿐만 아니라 특정 색인어의 사용범위를 기술해 줌으로써 색인어의 의미를 보다 명확히 정의해 주며 동시에 적절한 색인어의 선택이 가능하도록 한다.

시소러스의 목적은 크게 두 가지 측면에서 생각해 볼 수 있다. 하나는 색인자의 측면으로, 문헌의 주제를 나타내는 색인작업에 일관성을 유지하기 위한 것이고, 다른 하나는 탐색자의 측면으로, 문헌의 탐색을 용이하고 정확하게 하기 위한 것이다. 시소러스는 색인작업시 적절한 색인표목의 선택과 색인어의 통제를 위해 필요하며, 탐색시에는 적절한 탐색어의 선택이나 축소를 통해 검색효율을 조절하는데 사용된다.

시소러스가 정보의 저장과 탐색의 중요한 도구임에는 틀림없지만, 구성이 복잡하여 이용하기 어렵고, 한정된 분야에서만 이용이 가능하며, 보편화되어 있지도 않은 이유로 일반인에게 쉽게 이용되지는 않는다. 그러나 최근의 하이퍼텍스트 기술은 시소러스의 이용 가능성을 상당히 높여주고 있다. Shneiderman(1989)은 "수많은 조각들로 구성된 커다란 정보 덩어리가 있고, 이들 조각들이 서로 관계를 맺으며, 이용자들이 조그만 조각의 정보만을 원할 때 하이퍼텍스트는 유용하게 쓰인다"고 말하고 있다. 시소러스는 이러한 상황에 잘 들어맞는다고 본다. 전형적인 시소러스는 수많은 용어들을 포함하고, 이들은 서로 관계를 맺고 있으며, 시소러스 이용자들은 항상 부분적인 내용을 식별하는데 관심을 두기 때문이다.

본 연구는 시소러스의 대표적인 지침들과 하이퍼텍스트를 이용하여 문헌정보학분야 시소러스를 웹 상에 설계해 봄으로써 정보를 저장하고 탐색하는 사람들에게 발전된 주제 접근의 도구를 제공해 주는데 있다. 시소러스 구성요소간의 관계에 기초하여 본 연구에서 제시한 시소러스의 관계형 데이터베이스 모델은 모든 시소러스에 공통적으로 적용될 수 있을 것이다.

2. 연구의 배경

2. 1 시소러스 표준

시소러스의 표준에는 단일언어(mono-

lingual)와 다언어(multilingual)에 관한 것이 있지만 본 연구에서는 단일언어 시소러스의 표준으로 제한하였다. 단일언어 시소러스의 국제표준에는 국제표준기구(International Organization for Standardization)가 출판한 ISO 2788이 있으며, 이 표준은 1974년에 초판이, 그리고 1986년에 2판이 출판되었다. 단일언어 시소러스의 국가표준으로는 미국의 ANSI Z39.19, 영국의 BS 5723(ISO 2788과 동일), 프랑스의 AFNOR NFZ 47-100, 그리고 독일의 DIN 1463 등이 있다. 본 연구는 ISO의 단일언어 국제표준 시소러스와 미국과 영국의 단일언어 국가표준 시소러스를 참고하였다.

2. 1. 1 ISO 2788:1986

32쪽 분량의 이 시소러스 개발지침은 모두 10개의 장과 부록으로 구성되어 있다. 이중 제9장은 "용어와 관계의 배열"(display of terms and their relationships)을 다루며, 모두 세 가지의 기본 배열방법, 즉 알파벳순 배열, 체계적 배열, 도식 배열에 대하여 기술하고 있다(ISO 1986).

① 알파벳순 배열(alphabetical display)

알파벳순 배열에서는 디스크립터(descriptor; preferred term이라고도 한다)이든 비디스크립터(non-descriptor; non-preferred term이라고도 한다)이든 모든 용어들이 단일 알파벳 순서로 조직된다. 전통적인 알파벳순 배열은 1967년 미국에서 출판된 "공학 및 과학용어 시소러스"(Thesaurus of Engineering and Scientific Terms)로부터 시작되며, 이러한 유형의 시소러스는 단일계층 배열로 조직된다.

단일계층 배열은 시소러스 구축의 가장 기본적인 유형이지만 여기서는 계층을 구성하는 모든 상위개념어와 하위개념어를 파악할 수 없다. 여기서 발전한 형태가 복수계층 배열이다. 복수계층 배열에서는 계층을 들여쓰기에 의해 표시한다.

② 체계적 배열(systematic display)

체계적 배열은 두 부분으로 구성된다. 하나는 체계부분(systematic section)으로 용어의 논리적 상호관계에 따른 용어의 범주(categories) 또는 계층(hierarchies)을 말하며 여기에 용어들은 주제의 순서에 따라 배열된다. 다른 하나는 이용자를 체계부분의 적절한 곳으로 인도하는 알파벳순 색인부분(alphabetical index section)으로 체계부분은 알파벳순 색인부분에 의해 보완된다. 체계부분과 알파벳순 색인부분은 주소부호에 의해 연결된다. 주소부호는 체계부분에 배정되며, 알파벳순 색인부분에서는 참조로써 기능하게 된다. 알파벳순 색인부분은 위에서 이미 언급하였으므로 여기서는 체계부분만을 살펴본다.

체계부분의 조직유형은 다시 두 가지 유형으로 세분된다. 하나는 분야별 또는 학문분야별 조직(organization into fields or disciplines)이고 다른 하나는 패싯에 의한 조직(organization by facets)이다. 실제에 있어서 이 두 조직은 자주 결합하게 되는데, 시소러스는 우선 분야별로 조직되고, 다음에 각 분야 내의 개념을 조직하기 위해 패싯이 이용된다. 패싯은 세분되고, 계층 구성의 논리적 기

초를 제시하는 패싯 지시어(facet indicator)가 삽입될 수 있다. 용어는 세분된 패싯 안에서 알파벳순으로 정리된다.

③ 도식 배열(graphic display)

도식 배열이란 색인어와 그들의 관계를 2차원의 도식으로 배열하여 이용자가 적절한 색인어를 도표 상에서 선택할 수 있도록 한 것이다. ISO에서 언급한 도식 배열의 두 형태는 트리구조(tree structures)와 화살표그래프(arrowgraphs) 방식이다. 도식 배열은 위의 체계적 배열과 마찬가지로 알파벳순 색인부분을 포함한다.

도식 배열방식의 가장 큰 장점은 개념들의 환경을 한 눈에 볼 수 있다는 것이다. 도식 배열방식에서는 범위주기나 동등관계를 나타내지 않으며, 계층관계와 연관관계를 구별하지도 않는다. 모든 자세한 내용은 알파벳순 색인부분에서 지원 받게 된다.

도식 배열은 미국과 영국에서는 광범위하게 이용되지 아니한다. 한 연구에 의하면 1986년 현재 779개의 시소러스 중 단지 5.5%만이 도식 배열방식으로 되어있으며 이들의 대부분은 유럽의 국가들에서 만들어진 것으로 조사되고 있다(Bertrand-Gastaldy 1986).

2. 1. 2 ANSI/NISO Z39.19:1993

미국의 ANSI(American National Standards Institute)/NISO(National Information Standards Organization) Z39.19:1993은 단일언어 시소러스의 구축, 형식, 그리고 관리를 위한 미국의 국가 지침서이며 Z39.19:1980의 개정판이다(NISO 1994). 이 지침은 어떻게 디스크립터들을 나열하고, 어떻게 용어들간의 관계를 설정하며, 인쇄물이나 화면상에서 어떻게 정보를 나타낼 것인지를 알려주고 있다. 본 연구에서 주목한 부분은 제7장 "화면배열"(screen display)의 3절 "배열유형"(types of display)이며 여기서 제시하고 있는 주요 내용은 다음과 같다.

① 알파벳순 목록(alphabetical listing)

이용자가 검색어를 입력하였을 때, 이것이 디스크립터이든 기입어(entry term)이든 상관없이 시스템은 해당 용어의 앞뒤에 위치하는 용어들을 알파벳순으로 보여줄 수 있어야 한다. 여기서 기입어는 색인표목으로 채택되지 않은 용어이며 이러한 기입어는 색인어로 연결시켜 주어야 한다. 색인의 유용성은 기입어에 따라서 크게 좌우되며, 이용자가 검색에서 옳은 단어를 찾을 가능성을 한층 높여주게 된다. 데이터베이스에 연결된 시소러스는 디스크립터들의 게시(揭示)수도 표시할 수 있어야 한다.

② 순열배열(permuted displays)

순열색인(permuted index)은 다른 형태의 알파벳순 배열방식이다. 순열색인은 주로 KWIC(Key Word In Context) 혹은 KWOC(Key Word Out of Context)색인의 형태를 취하며, 용어들의 주요어를 접근점으로 삼아 나머지 전후 문장부분을 나열하거나(KWIC 색인의 경우), 주요어를 포함한 전체 문자열을 그 주요어 뒤에 나열하여 만든(KWOC 색인의 경우) 색인의 형태이다. 순열색인은 복합어를 구성하는 용어 중 색인어

가 아니기 때문에 단일용어로 나타나지 않는 어휘에 접근하는데 특히 효과적이다. 디스크립터와 비디스크립터에 대한 KWIC 혹은 KWOC색인은 시소러스의 보조색인으로 사용된다. 배열방식은 자순배열(letter by letter arrangement)과 어순배열(word by word arrangement)의 두 가지 방법 중에서 하나를 채택하며, 복합어의 순열목록은 모든 경우의 이용자에게 유리하다.

③ 계층과 분류(hierarchies and classification)

시소러스의 화면배열에서 다양한 수준의 계층표시가 가능하여야 하며, 계층에서의 들여쓰기는 필수적이다. 이용자는 계층의 여러 수준에서 디스크립터를 선택함으로써 검색의 확대와 축소가 가능해야 하며, 디스크립터에 일부 또는 전체의 상위개념어 또는 하위개념어도 첨가할 수 있어야 한다. 기호와 함께 쓰이는 트리구조(도식배열)는 이러한 탐색의 확장을 용이하게 하나 다른 구조로도 같은 효과를 얻을 수 있다. 예를 들어, 시스템이 각 디스크립터에 줄 번호(line number)를 부여하거나 하이퍼텍스트로 처리해 줌으로써 이용자는 검색을 확대시킬 수 있다. 계층구조에서 색인 또는 탐색에 이용되지 않는 패싯 지시어는 각괄호 등을 사용하여 디스크립터와 구별되어야 한다. 패싯 지시어에 역시 줄 번호가 부여될 수 있다.

④ 용어의 세부사항(term detail)

이용자는 어떤 목록에서나 용어를 선택할 수 있어야 하며, 전체든 부분이든 선택한 용어의 상세한 기술도 볼 수 있어야 한다. 이용자는 디스크립터의 변천, 범위주기, 정의뿐만 아니라 모든 용어의 동등, 계층, 연관관계와 해당 시소러스에서 만들어진 특별한 관계까지도 볼 수 있는 옵션을 가지고 있어야 한다.

⑤ 도식배열(graphic displays)

최근에 행해진 몇 개의 연구는 일부 이용자들에게 도식배열이 선형배열보다 더 효과적으로 개념사이의 관계를 나타내고 있음을 보여주고 있다. 시소러스를 도식배열 할 때에는 이용자의 범주와 검색습관을 고려하여야 한다.

본 연구에서는 ISO의 "용어와 관계의 배열"과 ANSI/NISO의 "화면배열유형"에서 선형배열만을 채택하였으며 도식배열에 대한 논의는 다음의 연구과제로 미루었다. 이는 선형배열을 도식배열과 함께 설계하는 것이 적합치 않다고 생각하였기 때문이다. 따라서 제5장의 하이퍼텍스트를 이용한 온라인 시소러스의 선형배열 설계는 ISO의 알파벳순 배열과 체계적 배열, ANSI/NISO의 알파벳순 목록, 순열배열, 계층과 분류, 용어의 세부사항을 모두 포함하는 인터페이스가 된다. 한편, ISO의 알파벳순 배열과 ANSI/NISO의 알파벳순 목록, 그리고 ISO의 체계적 배열과 ANSI/NISO의 계층과 분류내용은 많은 부분이 중복되고 있다. 결과적으로 본 연구의 제5장에서 설계한 내용은 알파벳순 목록(5.2), 순열색인(5.3), 분류와 계층(5.4), 용어의 세부사항(5.5) 등을 포함하게 된다.

2. 2 연구 과정

위에서 살펴본 ISO 2788:1986과 ANSI/NISO Z39.19:1993을 근거로, 제3장에서는 디스크립터와 UF(Used For)참조, BT(Broader Term)참조와 NT(Narrower Term)참조, RT(Related Term)참조와 RT참조의 관계와 이들의 연결(connectivity; 관계를 통해서 연결될 수 있는 개체들의 수)을 분석하였으며, 제4장에서는 이들을 관계형 데이터베이스에서 나타낼 수 있는 개념적 스키마와 논리적 스키마를 구상하였다. 개념적 스키마는 모든 이용자 관점을 통합한 전체 데이터베이스의 관점이고, 논리적 스키마는 하드웨어에 저장되는 데이터베이스의 물리적인 구조를 말한다. 개념적 스키마의 작성에는 개체-관계(E-R) 모델(entity-relationship model)이 적용되었다.

본 연구를 위해서 문헌정보학 분야 시소러스인 ASIS Thesaurus of Information Science and Librarianship(Milstead 1998)이 관계형 데이터베이스에 저장되었다. 테이블 형태로 표현되는 관계형 데이터베이스 모델은 그 구조의 간단명료함과 조작의 유연성으로 현재 널리 채택되고 있다. ASIS 시소러스는 1,300여 개의 디스크립터와 690여 개의 비디스크립터, 그리고 36개의 패싯 지시어로 구성되어 있다. 이들을 저장하기 위한 관계형 데이터베이스로는 Micorsoft사의 ACCESS 97이 이용되었다.

관련 레코드들의 조회, 삽입, 수정, 삭제, 결합을 위해서 관계형 데이터베이스의 질의어인 SQL(Structured Query Language)을 사용하였으며, 관계형 데이터베이스와 웹 연동을 위해서는 Windows NT운영체제 하에서 Microsoft사가 개발한 ASP(Active Server Pages)기술을 적용하였다.

3. 시소러스 구성요소간의 관계와 연결

기본적으로 모든 시소러스의 용어들 사이에는 동등관계(equivalent relationship), 계층관계(hierarchical relationship), 연관관계(associative relationship)가 성립한다. 이들의 관계를 디스크립터와 관련된 연결을 중심으로 살펴보면 다음과 같다.

3. 1 디스크립터와 UF참조

디스크립터와 UF참조는 동등관계를 갖는다. 동등관계는 색인작업 시 복수의 용어가 동일개념을 나타내는 경우 디스크립터와 비디스크립터간의 관계이다. 이때 USE참조는 디스크립터에 대한 접두기호로 사용되며, UF참조는 비디스크립터에 대한 접두기호로 사용된다. UF참조의 비디스크립터들은 USE참조에 의해 디스크립터로 연결된다.

하나의 디스크립터는 0개 또는 1개 이상의 UF참조를 가질 수 있다. 그리고 이론상 동등관계에서 UF참조는 하나의 디스크립터만을 가져야 하지만, 실제로는 하나의 UF참조는 여러 개의 디스크립터를 가지게 된다. 예를 들어, INSPEC 시소러스의 경우 800여 개의 UF참조는 하나 이상의 디스크립터를 가지고

있으며, 25개의 UF참조는 무려 5개 혹은 그 이상의 디스크립터를 가지는 것으로 조사되었다(Jones 1993). 디스크립터와 UF참조 사이에는 다:다(M:N)의 연결이 성립한다.

3. 2 BT/NT참조

BT/NT참조의 관계는 계층관계로 상위개념어와 하위개념어간의 관계를 말한다. 상위개념어는 전체를 나타내며, 하위개념어는 한 요소 또는 일부분을 나타낸다. BT참조는 디스크립터의 상위개념어를 나타내는 기호로 사용되며, NT참조는 디스크립터의 하위개념어를 나타내는 기호로 사용된다. 시소러스에서의 계층관계는 추상적인 범주보다는 실질적인 자연언어에 기초를 두고 있기 때문에 엄격한 계층구조를 이루지는 않는다. 즉 다중계층관계(poly-hierarchical relationships)에 의해서 몇몇 개념들은 하나 이상의 범주에 속할 수 있다. 예를 들어 "biochemistry"는 "biology"와 "chemistry"의 두 범주에 동시에 속할 수 있다. 따라서 하나의 NT참조는 여러 개의 BT참조를 가질 수 있으며, 결과적으로 NT참조와 BT참조 사이에는 다:다의 연결이 성립한다.

하나의 디스크립터는 0개 또는 하나 이상의 BT참조를 가지게 되며, 하나의 BT참조 역시 하나 또는 여러 개의 디스크립터를 가질 수 있다. 예를 들어 디스크립터 "automatic classification"은 "classification"과 "computer applications" 등의 BT참조를 가질 수 있으며, BT참조인 "computer application"은 "automatic classification", "automatic extracting", "automatic indexing" 등의 디스크립터를 가질 수도 있다. 디스크립터와 BT참조 사이에는 다:다의 연결이 성립한다. 디스크립터와 NT참조의 경우도 마찬가지로 다:다의 연결이 성립한다.

디스크립터, BT참조, NT참조간에는 순환관계도 성립한다. 예를 들어 "AA"라는 디스크립터가 "BB"라는 BT참조를 가질 때, "BB"가 디스크립터가 되면, "AA"는 "BB"의 NT참조가 되고, "AA"라는 디스크립터가 "CC"라는 NT참조를 가질 때, "CC"가 디스크립터가 되

information seeking
RT: information needs
information use

information needs
RT: information seeking
information use

information use
RT: information needs
information seeking

〈예 1〉 디스크립터와 RT참조의 다:다 연결

면, "AA"는 "CC"의 BT참조가 된다. 이는 다시 말하면 BT참조, NT참조 모두 디스크립터가 될 수 있음을 의미하기도 한다.

3. 3 RT/RT참조

RT참조는 동등관계나 계층관계는 아니지만 용어간에 개념적으로 상호 관련성을 갖는 연관관계이다. 연관관계에 있는 용어들은 색인작성과 탐색에 이용될 가능성이 있는 대체용어를 제시하는 것이 좋을 것이라는 심리적인 연상에 따라 만들어진 것이다.

하나의 디스크립터는 0개 또는 1개 이상의 RT참조를 가지며, RT참조는 순환하면서 디스크립터가 된다. 〈예 1〉은 하나의 디스크립터가 두 개의 RT참조를 가지고, 각각의 RT참조도 두 개의 디스크립터를 가지는 다:다의 연결을 보여준다. 디스크립터와 RT참조사이에는 다:다의 연결이 성립한다.

3. 4 기타

범위주기(Scope Note; SN)는 디스크립터를 사전의 내용과 같이 완전하게 정의해 주지는 않지만 사용범위를 기술해 줌으로써 이용자에게 정확한 디스크립터의 선정을 도와준다. 디스크립터는 0개 또는 1개의 SN은 가지게 되며, SN은 디스크립터에 종속하게 된다. 디스크립터와 SN 사이에는 1:1의 연결이 성립한다.

동형이의어(homograph)를 색인할 때에는 한정어(qualifier)를 부가하게 된다. 색인어를 한정어와 구별하기 위해서 주로 괄호 안에 용어를 추가하여 두 개 이상의 의미를 구별한다. 이러한 한정어는 디스크립터의 일부가 되기 때문에 디스크립터와 한정어 사이에는 연결이 없다.

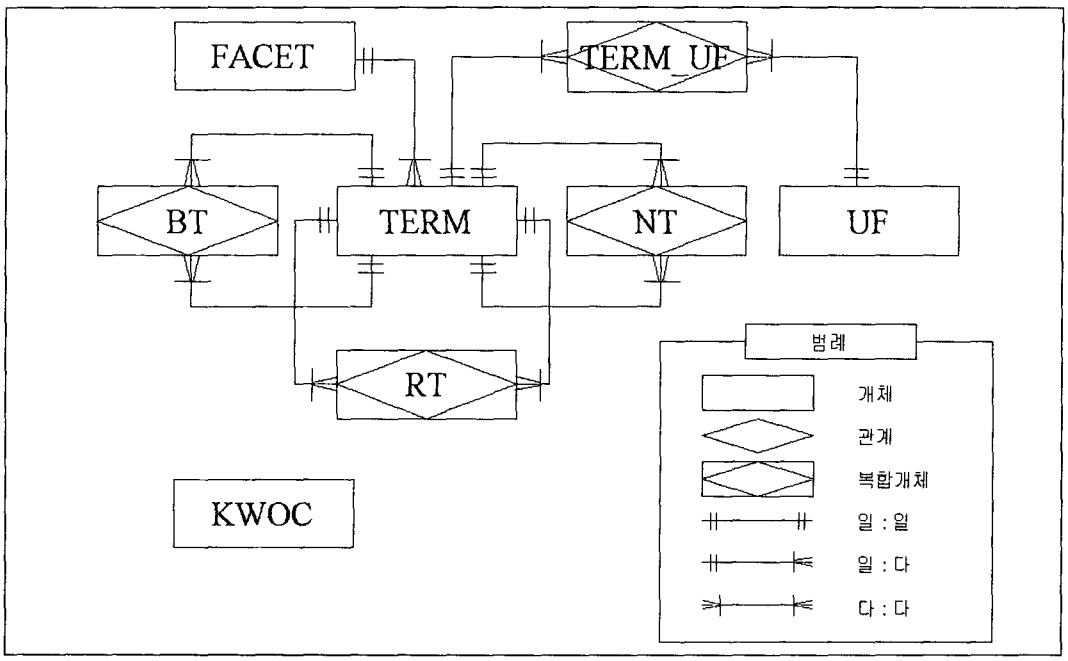
위에서 살펴본 시소러스의 용어와 이들의 모든 상호관계를 요약해 보면 〈표 1〉과 같다.

4. 시소러스 관계형 데이터베이스의 설계

관계형 데이터베이스 설계과정에서 가장 널리 사용되는 것으로 개체-관계(E-R) 모델이

〈표 1〉 시소러스 구성요소간의 관계와 연결

개체(entities)	관계(relationship)	연결(connectivity)
디스크립터 : UF참조	동등	다:다
BT참조 : NT참조	계층	다:다
디스크립터 : BT참조	계층	다:다
디스크립터 : NT참조	계층	다:다
RT참조 : RT참조	연관	다:다
디스크립터 : RT참조	연관	다:다
디스크립터 : SN(범위주기)	종속	1:1
디스크립터 : 한정어	부분	없음



〈그림 1〉 시소러스의 개체-관계 모델(개념적 스키마)

있다. E-R 모델은 데이터가 어떻게 하드웨어에 저장되는지 언급함이 없이, 개체와 그들의 관계성을 인간이 인지하는 방식과 유사한 형태로 기술할 수 있다는 장점을 가지고 있다. E-R 모델에서 묘사되는 기본적인 대상은 개체(entity)인데, 개체란 현실세계에서 스스로 존재할 수 있는 '어떤 것'이다. 개체는 어떤 사람, 어떤 물건, 어떤 건물 등과 같이 물리적으로 존재하는 유형의 것일 수도 있고, 어떤 학과목, 어떤 계약, 어떤 직업 등과 같이 개념적으로 존재하는 무형의 것일 수도 있다. 개체와 그들간의 관계를 나타내기 위해서는 직사각형이나 다이아몬드와 같은 특수 기호(그림 1의 범례참조)들이 사용된다. 〈그림 1〉은 〈표 1〉 시소러스 구성요소간의 관계와 연결을 E-R 모델로 나타낸 것이며, 이를 요약하여 설명하면 다음과 같다.

앞에서 살펴본 바와 같이 시소러스를 구성하는 요소 중 BT/NT/RT참조의 용어들은 모두 디스크립터가 될 수 있으며, BT/NT/RT 참조의 용어들은 디스크립터와 각각 다:다의 관계를 갖는다. 관계형 모델에서 다:다의 경우는 일:다(1:M)의 관계로 전환되어야 하며, 이를 위해서는 다:다의 관계를 보여줄 수 있는 다리(bridge)를 필요로 한다. 이 다리는 연결된 개체형의 주 키(primary key)들로 구성되는 새로운 개체이며 이러한 개체를 복합개체(composite entities)라 한다(Rob and Coronel 1993).

〈그림 1〉의 TERM개체는 디스크립터들의 집합이며 BT, NT, RT는 각각 개체이면서 TERM개체와 관계를 갖는 복합개체가 된다.

〈표 2〉 시소러스의 논리적 스키마

테 이 블	속 성
TERM	TERM_ID(디스크립터 ID)
	FACET_ID(패싯 ID)
	SN(범위주기)
	TERM_NAME(디스크립터 이름)
	COUNT(게재 수를 위한 수)
BT	TERM_ID(디스크립터 ID)
	BT_ID(BT를 나타내는 디스크립터 ID)
NT	TERM_ID(디스크립터 ID)
	NT_ID(NT를 나타내는 디스크립터 ID)
RT	TERM_ID(디스크립터 ID)
	RT_ID(RT를 나타내는 디스크립터 ID)
UF	UF_ID(비디스크립터 ID)
	UF_NAME(비디스크립터 이름)
TERM_UF	TERM_ID(디스크립터 ID)
	UF_ID(비디스크립터 ID)
FACET	FACET_ID(패싯 ID)
	FACET_NAME(패싯 이름)
	FACET-INDICATOR(패싯 지시어)
KWOC	WORD(단어의 집합)

복합개체 BT, NT, RT는 또한 TERM개체와 서로 순환하면서 발생하는 순환개체(recursive entity)가 된다.

시소러스에서 UF참조는 디스크립터가 될 수 없으며, 따라서 별개의 UF개체를 필요로 한다. UF개체와 디스크립터 사이에는 역시 다:다의 관계를 가지며, 이들 사이에는 다리 역할을 해주는 TERM_UF라는 복합개체를 필요로 한다.

ISO와 ANSI/NISO의 배열표준에는 포함되지 않지만, 본 연구의 목적을 위해서 FACET개체를 만들었다. 시소러스에서 패싯은 논리적인 기초를 지시하기 위해서 패싯 지시어를 가지며 이것 자체는 색인어로 쓰이지 않는다. 각각의 패싯 지시어 아래에는 디스크립터들이 존재하게 되며 이들을 나타내기 위해서는 FACET이라는 개체를 필요로 한다. 하나의 패싯 지시어는 여러 개의 디스크립터를 가질 수 있지만, 하나의 디스크립터는 하나의 패싯 지시어에만 속할 수 있다. 따라서 FACET개체와 디스크립터 사이에는 일:다의 연결관계가 성립한다.

KWOC개체는 KWOC색인을 위한 것으로 디스크립터와 비디스크립터를 구성하는 모든 단어들이 저장된다. KWOC개체와 TERM개체, KWOC개체와 UF개체 사이에는 각각 의존관계와 다:다의 연결이 성립한다. 예를 들어, TERM개체의 개체 예(instance)인

"access points" 와 UF개체의 개체 예인 "access vocabularies"는 KWOC개체에서 "access", "points", "vocabularies"로 표현되며(의존관계), KWOC개체 예인 "access"는 "access control", "access points", "access to resources" 등을 가질 수 있다(다:다의 연결). 〈그림 1〉에 대한 논리적 스키마는 〈표 2〉와 같다.

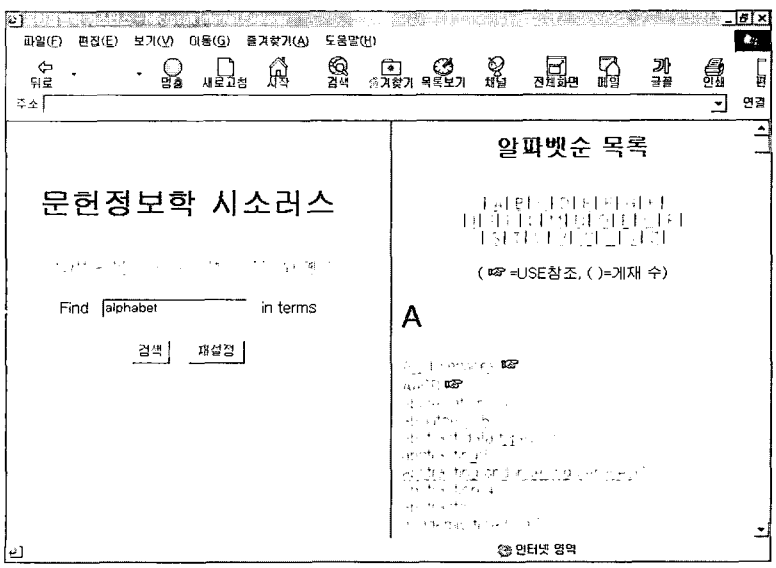
5. 온라인 시소러스의 설계

본 연구의 하이퍼텍스트를 이용한 문헌정보학 분야 온라인 시소러스의 선형배열 설계는 크게 4개 부분으로 나누어진다. 검색창에 의한 접근(5.1), 알파벳순 목록에 의한 접근(5.2), KWOC색인에 의한 접근(5.3), 패싯 분류와 계층에 의한 접근(5.4)이 그것이다. 〈그림 2〉의 좌측화면은 문헌정보학 시소러스 브라우저의 첫 화면으로 이들 네 가지의 접근 방법을 보여주고 있다.

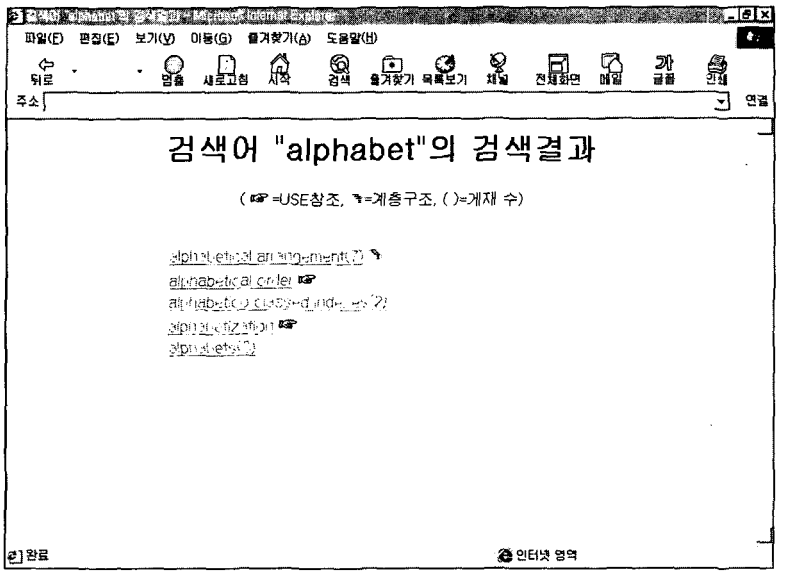
5. 1 검색 창에 의한 접근

검색 창에 의한 접근은 순열색인의 KWIC 색인과 매우 유사하다. 검색어의 문자 또는 단어가 〈그림 1〉의 TERM개체나 UF개체 안에 있으면 모두 검색되기 때문이다. 예를 들어 검색 창에 "a"를 입력하면 "a"문자가 들어있는 모든 용어들이 검색되어진다.

〈그림 3〉은 〈그림 2〉의 검색 창에 검색어 "alphabet"을 입력했을 때 얻은 결과 값들을 보여준다. "alphabet"이라는 문자열이 포함되어 있으면 모두 검색되었다. 검색결과에는 또한 USE참조, 계층구조, 게재 수의 세 가지 옵션도 있음을 알 수 있다. 계층구조에 대한 설



〈그림 2〉 첫 화면



〈그림 3〉 검색 창에 의한 검색결과

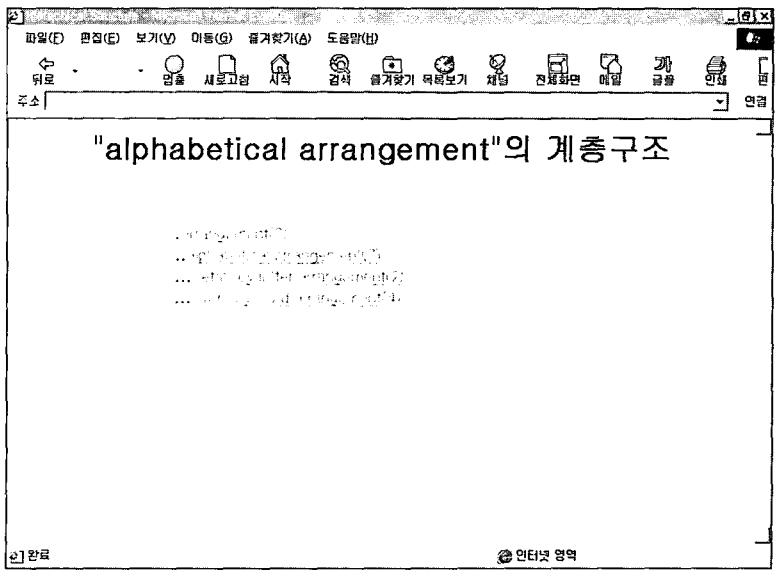
명은 다음과 같다.

ANSI/NISO의 지침에 의하면 이용자는 계층의 여러 수준에서 디스크립터를 선택함으로써 검색의 확대와 축소가 가능해야 하며, 디스크립터에 일부 또는 전체의 상위개념어 또는 하위개념어도 첨가할 수 있어야 한다고 말하고 있다. 본 연구에서 계층구조는 특정 디스크립터가 BT/NT관계를 가지게 되면 ▸표시를 자동으로 나타내게 했으며, ▸를 클릭하면 디스크립터가 어느 계층에 위치하고 있는지를 보여 주게 된다. 계층의 단계는 "•"로 나타내었다.

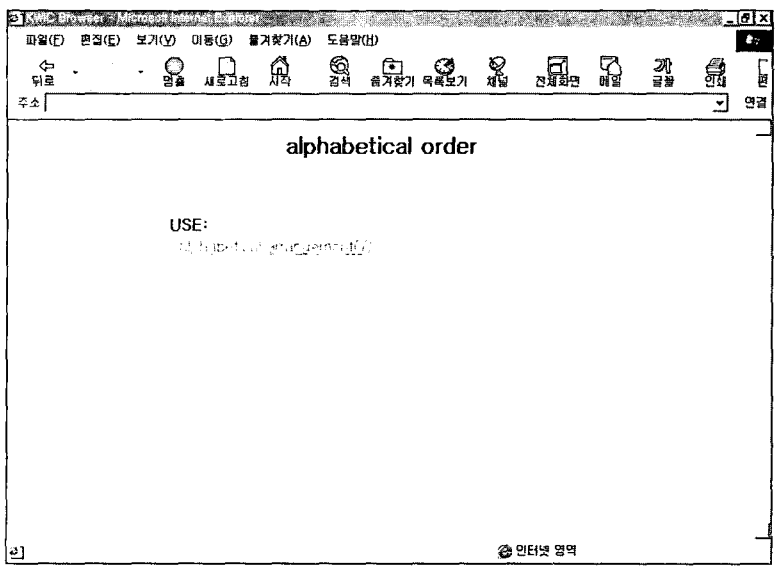
〈그림 3〉 "alphabetical arrangement"의 클릭 예는 〈그림 4〉와 같다. 〈그림 4〉에서 "• •" 표시의 "alphabetical arrangement"는 "•" 표시의 상위개념어(BT참조) "arrangement"를 가지며, "• • •"표시의 하위개념어(NT참조) "letter by letter arrangement" 와 "word by word arrangement"를 가진다. 만약 "letter by letter arrangement" 또는 "word by word arrangement"가 하위개념어를 가진다면 "• • • •"표시로 이들 밑에 나열 될 것이다. 모든 계층관계는 하이퍼텍스트로 연결되어 있으며 어디서나 검색의 확대와 축소가 가능하다(5.5 참조). USE참조, 게재 수는 아래 5.2에서 설명하였다.

5. 2 알파벳순 목록

〈그림 2〉 첫 화면의 우측 화면은 알파벳순 목록에 의한 접근을 보여주며, 모든 용어(기입어 포함)들은 알파벳순으로 어순배열로 나열된다. ANSI/NISO의 지침에 의해 USE참조는 ☞ 표시로 디스크립터와 구별하였다. 〈그림 5〉의 USE참조 표시는 〈그림 3〉의



〈그림 4〉 검색 창에 의한 검색어의 계층구조



〈그림 5〉 USE참조 표시

"alphabetical order" 클릭 예를 보여준다.

디스크립터와 따라다니는 괄호는 게재 수를 의미한다. 가령 "alphabetical arrangement(7)"에서 7은 디스크립터 "alphabetical arrangement"가 시소러스 내에서 7번 출현하였음을 나타낸다. 게재 수는 해당 디스크립터의 시소러스 내에서의 출현빈도를 의미하며, 해당 디스크립터의 게재 수가 높으면,

이는 시소러스 내에서 자주 참조되는 용어임을 알 수 있다. 알파벳순 목록에서 용어들은 A-Z목록을 통하여 쉽게 접근할 수 있다.

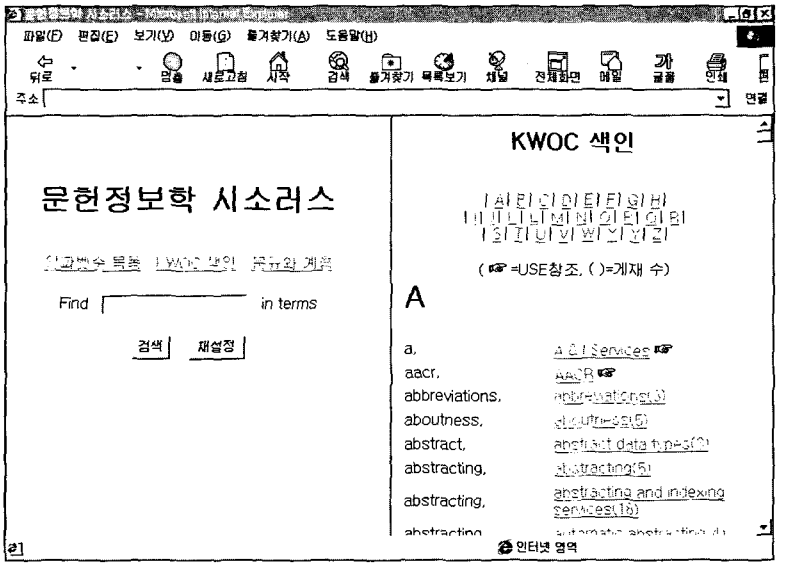
5. 3 KWOC색인

KWOC색인은 주요어가 접근점이 되어 주로 왼쪽에 위치하고, 전체 문자열이 그 뒤에 나열되는 순열색인의 한 형태이다. KWOC색인에 의한 접근에서는 모든 디스크립터와 기입어들이 KWOC색인으로 배열된다. 기입어를 디스크립터와 구별하기 위해서 역시 ☞ 기호를 사용하였으며, 각 디스크립터의 뒤에 붙은 괄호 안의 숫자는 게재 수를 의미한다.

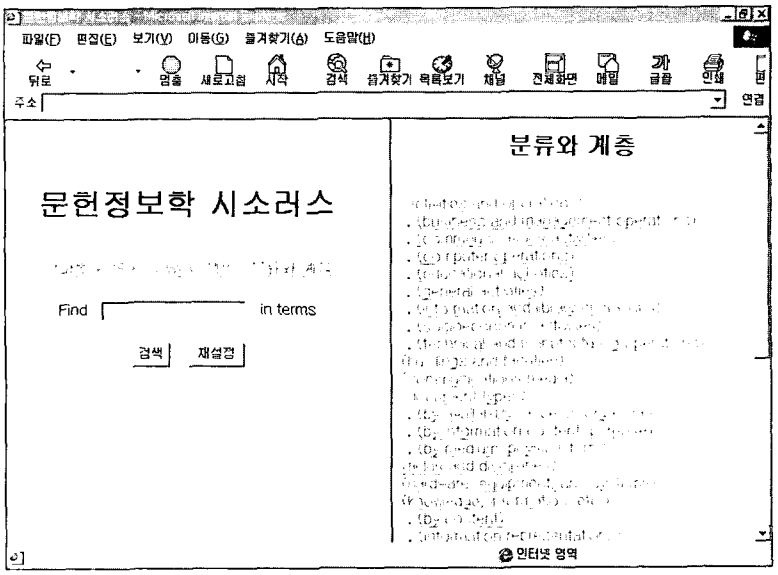
디스크립터나 기입어의 동형이의어(homographs)를 구별하기 위해 괄호 안에 포함되는 수식어(qualifier)는 디스크립터의 일부로 간주하여 KWOC색인에 포함하였다. 그러나 색인과 검색 시 무시되는 전치사, 관사, 접속사 등의 불용어(stopwords)는 KWOC색인에 포함되지 않았다. KWOC색인은 어순배열로 나열하였으며 대·소문자는 구별하지 않았다. 알파벳순 목록과 마찬가지로 모든 용어는 A-Z목록을 통하여 쉽게 접근할 수 있다. 〈그림 6〉의 우측화면은 KWOC색인의 예를 보여주고 있다.

5. 4 패싯분류와 계층

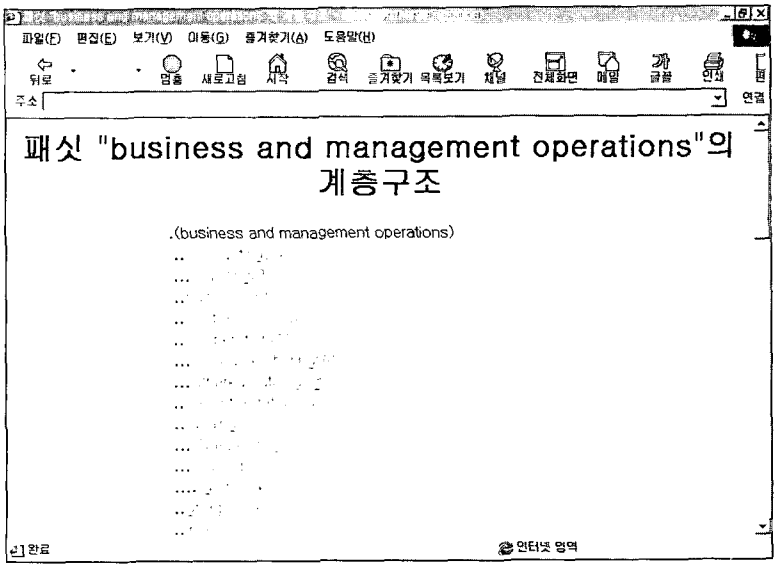
패싯분류와 계층에 의한 접근방법은 이용자가 광범위한 주제를 마음속에 가지고 있을 때 주로 이용된다. 〈그림 7〉의 우측화면은 36개의 패싯 지시어중 일부를 보여주고 있으며, 패싯 지시어는 괄호로 표현하였다. 패싯간의 계층은 검색어의 계층구조와 마찬가지로 "•"를 사용하였다.



〈그림 6〉 KWOC색인에 의한 접근

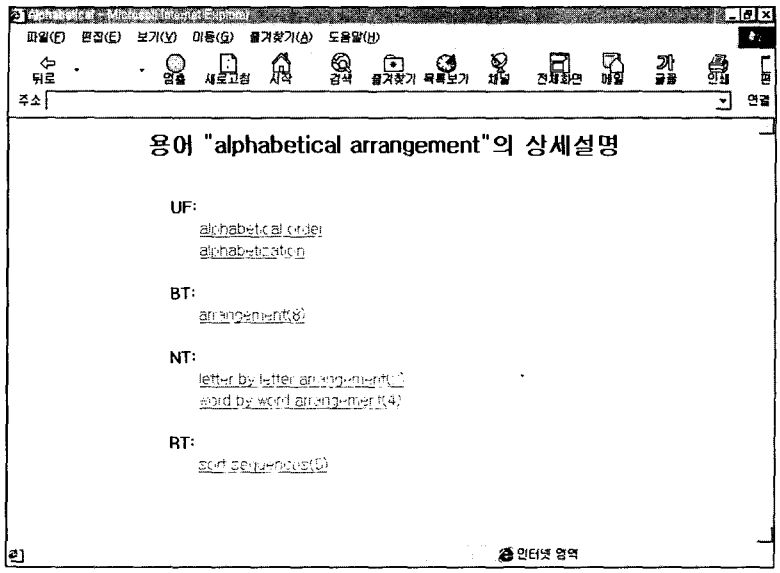


〈그림 7〉 패싯분류의 표시



〈그림 8〉 패싯의 계층구조

〈그림 8〉은 패싯분류의 계층구조를 보여주며 "•"이 증가할 때마다 하위개념을 나타낸다. 예를 들어, 패싯 지시어 (activities and operations)는 "•"로 표시된 하위 패싯 지시어 (business and management operations)를 가지며, 이는 "• •"로 표시된 "accounting", "business", "funding"과 같은 하위개념어들을 가진다. 그리고



〈그림 9〉 용어의 상세설명 예

"accounting"은 "• • •"로 표시된 하위개념어 "auditing"을 가진다.

5. 5 용어의 세부사항

〈그림 2〉의 첫 화면에서부터 〈그림 8〉 패싯의 계층구조까지 모든 내용은 하이퍼텍스트로 연결되어 있다. 어느 단계에서도 용어를 선택할 수 있으며, 해당 용어에 대한 상세한 기술도 볼 수 있다. 해당 용어에는 디스크립터의 범위주기 뿐만 아니라 모든 용어의 동등, 계층, 연관관계를 포함하게 된다. 〈그림 9〉는 모든 화면에서 접근 가능한 용어의 상세설명 예를 보여주고 있다.

6. 결론 및 제언

전문탐색자들은 시소러스에 많은 부분을 의존한다(Fidel 1991). 그러나 전문탐색자들이 실지로 어떻게 시소러스를 이용하고 있는지에 대하여는 별로 알려진 것이 없다. 예를 들면, 탐색자들은 어떤 시소러스의 배열을 이용하는가? 정보의 탐색자들이 시소러스를 이용하여 원하는 디스크립터 또는 비디스크립터를 찾는데 얼마나 많은 관계추적을 거쳐야 하는가? 등이다. 이러한 질문의 답은 인쇄형 시소러스를 가지고는 대답하기 어렵다. 그러나 하이퍼텍스트를 이용한 시소러스는 시스템과 더불어 이용자의 반응을 기록함으로써 위의 질문에 대한 답을 어느 정도 가능케 한다.

본 연구에서는 ISO와 ANSI/NISO 시소러스의 작성지침에서 필요한 부분을 참고하여 문헌정보학분야 시소러스를 하이퍼텍스트를

이용하여 선형배열로 웹 상에 설계해 보았다. 이를 위하여 시소러스 작성지침에 나타난 화면배열유형, 시소러스 구성요소간의 관계, 시소러스 구성요소간의 연결, 시소러스의 개체-관계(E-R) 모델, 시소러스의 논리적 스키마 등이 논의되었다. 하이퍼텍스트를 이용한 온라인 시소러스는 정보를 저장하고, 탐색하는 사람들에게 발전된 주제접근의 도구를 제공해 줄 수 있을 것이며, 이용자의 시소러스 탐색 유형을 연구하는 기초가 될 수 있을 것이라 기대한다.

앞으로 좀 더 나은 온라인 시소러스의 디자인과 유용성 테스트에 대한 많은 과제가 남아 있다. 온라인 시소러스에는 어떠한 특징들을 포함하여야 하는가? 다양한 용어관계들은 어떻게 표현되어야 하는가? 온라인 시소러스와 인쇄형 시소러스 중에서 이용자들은 어느 쪽을 더 선호하는가? 이러한 질문에 답하기 위해서는 많은 연구들이 수행되어야 할 것이다. 마지막으로 우리 나라에도 다양한 시소러스가 만들어지고, 이들이 웹 상에 등장하는 시기를 기대해 본다.

감사의 글: 본 연구에 도움을 주신 성균관대학교 문헌정보학과 오삼균 교수와 프로그래밍을 도와준 김덕성 조교에게 고마움을 표시합니다.

참 고 문 헌

Bertrand-Gastaldy, S. 1986. "Improved Design of Graphic Displays in Thesauri-through Technology and Egonomics." *Journal of Documentation.* 42(4): 225-251.

Fidel, R. 1991. "Searchers Selection of Search Keys: II. Controlled Vocabulary or Free-Text Searching." *Journal of the American Society for Information Science.* 42(7): 501-514.

ISO. 1986. *Documentation-Guidelines for the Establishment and Development of Monolingual Thesauri.* Geneva, Switzerland: International Organization for Standardization. (ISO 2788:1986).

Jones, S. 1993. "A Thesaurus Data Model for an Intelligent Retrieval System." *Journal of Information Science.* 19: 167-176.

Milstead, J. 1998. *ASIS Thesaurus of Information Science and Librarianship.* 2nd ed. Medford, NJ: Information Today, Inc.

NISO. 1994. *Guidelines for the Construction, Format, and Management of Monolingual Thesauri.* Bethesda, MD: National Information Standards Organization Press. (ANSI/NISO Z39.19:1993).

Rob, P. and Carlos C. 1993. Database Systems: *Design, Implementation, and Management.* Belmont, California: Wadsworth Publishing Company.

Shneiderman, B. 1989. "The Society of Text." In *Reflections on Authoring, Editing, and Managing Hypertext.* E. Barrett(Ed.). Cambridge, MA: MIT Press.