다국어 통합 개념체계의 구축 방안 연구*

Development of a Multilingual Integrated Concept System

정영미(Young Mee Chung)**, 김명옥(Myoung Ock Kim)***, 이재윤(Jae Yun Lee)****,
한승희(Seunghee Han)*****, 유재복(Jae Bok Yoo)******

초 록

과학기술 분류표, 시소러스, 용어사전 등의 주요한 색인 및 검색 도구를 한국어, 영어, 일본어의 3개 언어로 통합 구축하고 활용할 수 있도록 다기능, 다국어 과학기술 통합 개념체계의 개발 방안을 마련하였다. 개념을 기본 단위로 시소러스 모델을 개발하였으며, 용어사전 레코드는 ISO 12620 표준에 근거하여 필수요소를 지정하였다. 또한 과학기술분야 표준분류표를 대분류 수준까지 작성하고 기존 분류표와의 매핑 테이블을 작성하여 다른 분류표를 통한 접근이 가능하도록 하였다. 시소러스, 용어사전, 분류표의 원활한 상호 연계와 운용을 위해서 통합 개념체계 모형을 설계하였다. 본 연구에서 개발한 통합 개념체계를 이용하여 원자력 분야를 대상으로 한 프로토타입 시스템을 구축하고 실제 검색 사례를 제시하였다.

ABSTRACT

키워드 : 통합 개념체계, 지식베이스, 시소러스, 분류표, 용어사전, 다국어 시소러스, 패싯
integrated concept system, knowledge-base, thesaurus, classification system, terminology, terminological database, multilingual thesaurus, facet

---

* 이 논문은 2001년도 한국과학기술정보연구원 기본연구사업에 의하여 지원되었음.
** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)
*** 숭의여자대학교 문헌정보과 교수(kimm@sewc.ac.kr)
**** 연세대학교 문헌정보학과 시간강사(memexlee@lis.yonsei.ac.kr)
***** 서울여자대학교 문헌정보학과 시간강사(libinfo@lis.yonsei.ac.kr)
****** 한국원자력연구소 기술정보실 선임연구원(jbyoo@kaeri.re.kr)

1. 서론

신속하고 정확한 정보 유통을 위해서는 수집된 정보를 체계적으로 조직하여 이용시킬 필요가 있는데, 이때 적용되는 기법이 분류, 색인, 검색, 필터링 등이다. 특히 최근에는 언어의 장벽을 넘어서 다국어 정보를 처리하기 위하여 자동번역과 교차언어 정보검색에 대한 연구가 활발히 진행되고 있다.

이와 같은 정보 처리 및 유통 환경을 구축하기 위해서는 시소러스나 용어사전과 같은 개념 지식베이스의 활용이 필수적이다. 시소러스와 용어사전은 원래 개별적으로 발전해왔으나, 최근에는 구축 작업의 경제성과 일관성을 위해서 통합하여 개발할 필요성이 제기되고 있다(Calzolari 1988). 이미 1990년대 들어 여러 학자들이 시소러스를 확장한 통합 개념 지식베이스의 필요성을 지속적으로 제기한 바가 있다(Schmitz-Esser 1990; 1991; 1999; Soergel 1996; 1999).

국내에서 개발된 시소러스는 수가 많지 않으며 실제 사용도 활발하지 않은 실정이다. 반면에 외국의 경우 여러 시소러스가 개발되고 꾸준히 개정되면서 활발히 이용되고 있으며, 웹 환경의 온라인 정보서비스에서도 대부분 시소러스에 의한 검색 기능을 제공하고 있다(Shiri and Revie 2000). 특히 다국어 처리를 위해서 주제분야별 다국어 시소러스나, EuroWordNet과 같은 일반 어휘의 다국어 시소러스의 구축이 활발히 진행되고 있다.

국내의 시소러스 구축 및 활용이 활발하지 못한 것은 정보 분석 업무의 위축이 한 원인이긴 하지만, 전문 데이터베이스의 활성화와 다국어 정보의 유통으로 인해 급속하게 변화한 실제 응용 영역을 고려하지 못한 시스템 설계 과정에도 문제가 있는 것으로 보인다. 적어도 이미 구축된 시소러스나 용어사전을 활용함으로써 생산성을 높이고 기존 도구의 개정 작업을 지속적으로 유지할 수 있는 체계가 확립되어야 할 것이다.

한편 동일한 주제분석 도구라는 측면에서 시소러스와 분류표를 통합할 필요성도 오래전부터 제기되어왔다. 특히 통합체계에 대한 필요성은 대부분의 정보서비스가 인터넷을 통해서 제공되면서 더욱 커졌다. 국회도서관과 같은 기존 도서관의 서지 데이터베이스, LG 상남과 같은 디지털도서관 서비스, KISTI나 KERIS와 같은 기관의 연구정보 서비스, Dialog나 OVID와 같은 온라인 데이터베이스 서비스, PQD나 IDEAL과 같은 전자저널 서비스 등이 모두 인터넷에서 웹을 통해 제공되고 있다. 복수의 정보서비스가 인터넷상의 분산 데이터베이스인 것처럼 제공되고 이용되는 것이다. 따라서 각 서비스에서 사용되는 별개의 주제접근 도구의 차이가 이제는 심각한 문제로 부각되어 통합체계에 대한 필요성이 더욱 절실하게 되었다. 영국의 HILT(High-Level Thesaurus) 프로젝트에서 최근 조사한 바에 따르면 조사대상 기관 중 83%가 기관간 교차 주제탐색의 필요성을 느끼고 있으며 62%의 기관이 해결책을 심각하게 고민하고 있는 것으로 나타나고 있다(Wake and Nicholson 2001). 이런 통합체계에 대한 최근의 연구에서는 기본 엔트리로 단어나 용어가 아닌 개념을 이용하려는 경향이 나타나고 있다(Seorgel 1996).

국제적인 학술정보 유통의 활성화를 통한 국내 과학기술의 발전을 위해서는 과학기술 분야를 포괄하는 표준 분류표, 시소러스, 용어사전의 구축이 필요하며, 이들의 효율적인 개발 및 유지, 그리고 이용을 위해서는 개념을 기본 엔트리로 하는 통합체계를 우선적으로 개발해야 할 것이다.

통합 개념체계의 선행 사례로는 개념을 중심으로 구축된 WordNet, EDR 전자사전, UMLS 를 들 수가 있다. WordNet은 인간의 어휘지식에 대한 심리언어학 연구의 성과를 토대로 1985년부터 프린스턴 대학 인지과학연구실이 구축해 온 영어어휘 데이터베이스이다(Miller et al. 1990). EDR 전자사전은 일본의 전자사전연구소에서 9년 간의 국가 지원 프로젝트에 의해 개발한 것이다(Japan Electronic Dictionary Research Institute 1988). UMLS(Unified Medical Language System)는 미국 국립의학도서관(NLM)이 의료 및 보건 전문가와 연구자들이 다양한 정보원의 정보를 통합하여 검색할 수 있게 하고 특허 기록, 서지 데이터베이스, 사실 데이터베이스, 전문가 시스템 등의 상이한 정보시스템간의 연결을 용이하게 할 수 있도록 1986년부터 개발하여

지속적으로 갱신하고 있는 체계이다(National Library of Medicine 2001).

이들은 개념을 기본 단위로 하고 있으므로 다국어 처리도 개념을 통해 연결하고 있다. EUROVOC 시소러스, UNBIS 시소러스, GEMET과 같은 기존의 다국어 시소러스들이 대역어간의 직접 매핑을 통해서 다국어를 처리하는 것과는 조금 다른 방법이다. WordNet의 다국어판인 EuroWordNet은 언어중립적인 개념색인 목록인 ILI(Inter-Lingual-Index)를 통하여 여러 언어를 연결하고 있다. UMLS에서는 개념 식별기호와 용어 식별기호, 문자열 식별기호를 별도로 할당하고 각 언어의 용어마다 부여된 용어 식별기호를 동일 개념으로 연결하는 것으로 대응어를 처리하고 있다.

용어사전과 시소러스는 정보검색이나 자연언어처리를 위한 지식베이스로서도 활발히 연구되고 있다. 이들 연구에서는 각 용어에 대해서 패싯이나 속성을 부여할 필요를 제기하고 있다(박영자, 송만석 1992; 윤광진 1989). 기존 시소러스에서도 패싯을 도입한 경우가 다수 있다. 사물, 장치, 부분, 속성, 과정, 응용, 효과의 기본패싯을 사용한 Thesaurofacet, 11개의 역할기호를 사용한 EJT 시소러스, 8개의 패싯을 사용한 CIT 시소러스, 616종의 패싯지시자를 사용한 ROOT 시소러스 등을 예로 들 수가 있다.

그런데 지식베이스에 대한 패싯 적용 연구는 대부분 용어나 개념 자체에 패싯을 할당하는 문제를 다루는 반면에 시소러스에 대한 패싯 적용 연구는 주로 개념간의 관계를 구분할 때 패싯을 사용하는 문제를 다룬다는 차이가 있다. UMLS의 경우에는 이 두 가지를 함께 적용하였다. UMLS 의 일부인 의미망에서는 개념간의 관계를 5가지 패싯(물질, 공간, 기능, 시간, 개념)을 기준으로 총 51 가지의 관계를 정의하였고, 개념 자체도 개체와 사건의 두 집합으로 나누어 최대 7계층까지에 이르는 개념 속성 분류를 적용하고 있다.

분류표와 시소러스가 주제나 개념을 지향하는 반면에 용어사전은 용어 표기를 비롯한 용어 자체에 대한 정보를 위주로 구성된다. 용어사전 및 용어 데이터베이스의 형식 표준은 주로 교환을 위한 목적으로 개발되어왔다. 대표적인 것이 ISO 12200으로 지정된 MARTIF 표준(International Organization for Standardization 1999)이며, TEI P4 지침(Text Encoding Initiative Consortium 2001)의 E-TIF도 같은 목적에서 개발되었다.

선행 연구 및 사례를 검토한 결과 본 연구에서는 과학기술정보의 색인 구축과 검색 성능의 극대화를 위하여 개념을 단위로 하는 색인 및 검색을 위한 통합적인 도구를 개발하고자 한다. 즉, 과학기술 분야의 분류표, 한·영·일 대역 시소러스, 용어사전을 개발하여 통합 개념체계로 연계함으로써 주제 계층과 개념들간의 관계를 반영한 색인 데이터베이스를 제작 및 관리하게 하며, 검색 시 이용자들이 디스크립터, 분류기호, 관련 용어의 한, 영, 일 형태 중 어느 접근점이라도 사용할 수 있게 함으로써 검색의 편리성과 효율성을 증대시키고자 하는 것이다.

2. 통합 개념체계의 구성

통합 개념체계를 설계하면서 고려한 기본 원칙은 다음과 같다.

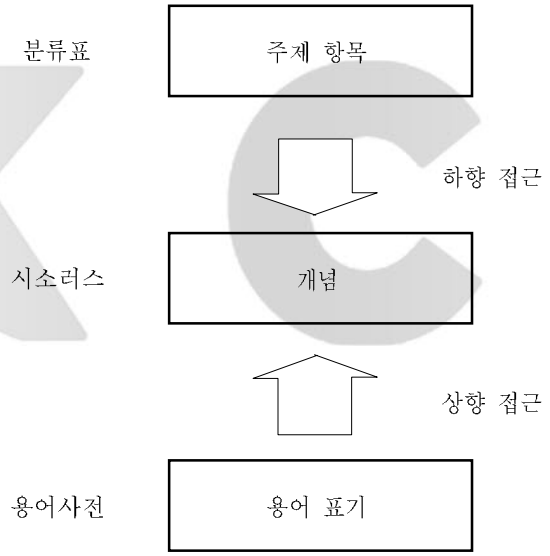
① 개념을 기본단위로 구성하였다. 특정 용어나 문자열이 아닌 개념을 기본단위로 함으로써 용어들의 연결, 분류주제를 통한 용어로의 접근, 용어 정보의 관리를 용이하게 하였다.
② 한국어를 중심언어로 선택하였다. 한국어, 영어, 일어의 다국어를 고려하는 체계이긴 하지만, 사용자를 한국어를 모국어로 하는 사람으로 전제하여 개념 표현이나 용어사전 설계에 반영하였다.
③ 패싯(속성)을 고려하였다. 개념 정보의 하나로 각 개념의 패싯유형을 지정하여 개념 정보와 관계 정보를 보완하였다.
④ 모듈화를 추구하였다. 분류표, 시소러스, 용어사전의 복수 체계를 수용하므로 동일 정보의 중복을 방

지하고 다양한 응용 시스템에 최적화해서 적용할 수 있도록 구성 요소를 최대한 모듈화하였다.

⑤ 개방성을 추구하였다. 다언어, 복수 체계를 통한 주제 접근에서 핵심 요소로 기능할 수 있도록 다른 언어나 체계의 수용 및 연결이 용이하게끔 개방적으로 설계하였다.

통합 개념체계의 기본 단위는 개념과 용어가 된다. 여기서는 개념을 의미적으로 동일하거나 화용론적으로 같은 뜻을 표현하는데 사용되는 용어의 집합으로 간주하였다. 사전적인 의미로는 비록 약간 차이가 있는 두 용어라 할 지라도 특정 분야에서 전문용어로 사용될 때 같은 개념을 나타낸다면 특정 맥락에서 화용론적인 동의어 관계로 간주하였다. 즉 시소러스에서 우선어와 비우선어는 모두 동일 개념을 나타내는 것으로 생각하였다.

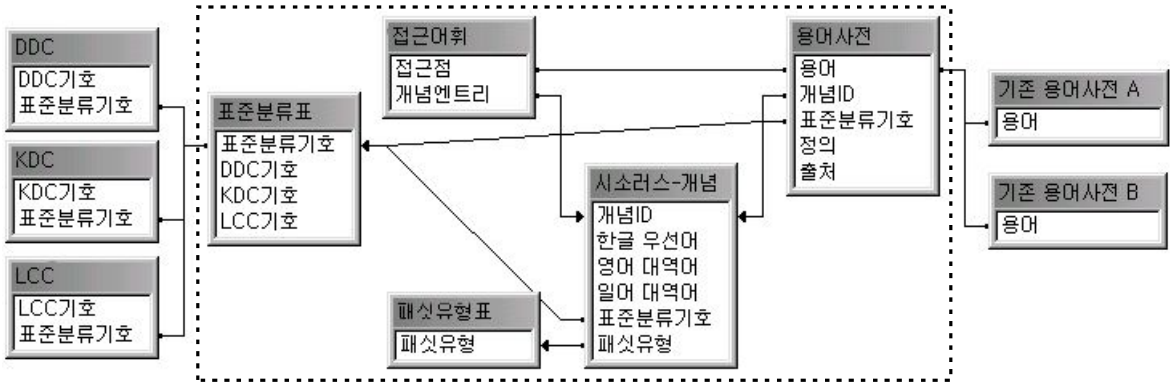
전체적으로 통합 개념체계는 시소러스, 분류표, 용어사전의 세 부분으로 구성된다. 시소러스는 개념을 표현하는 용어를 지시하고 개념간의 관계를 알려준다. 분류표는 계층적 주제분류를 통해 시소러스에 포함된 각 개념의 주제 분야를 나타낸다. 용어사전은 시소러스에서 개념을 표현하는 데 사용되는 각 용어에 대해서 개념정보, 문법정보, 관리정보를 저장한다. 통합 개념체계의 핵심 구성 단위인 개념에 대한 접근은 그림 1과 같이 분류표를 통해서 주제 항목에서 하향으로 접근할 수가 있으며, 용어사전에 등록된 실제 용어 표기를 통해서 상향으로 접근할 수도 있다.



<그림 1> 통합 개념체계에서 개념에로의 접근

분류표의 기본 단위는 주제, 시소러스의 기본 단위는 개념, 용어사전의 기본 단위는 용어가 된다. 범위 측면에서는 분류표가 다루는 주제가 가장 넓은 단위인 반면 용어사전이 다루는 용어가 가장 좁은 의미로 한정된 단위라고 볼 수 있다. 실제 사용에 있어서도 분류표는 대개 큰 주제로부터 세부 주제로 좁혀 들어가는 방식으로 접근하게 되는 반면에, 용어사전은 가장 구체적인 단위인 용어 자체로 직접 접근하여 이용하게 된다. 시소러스는 분류표와 용어사전의 중간 수준에서 포괄적인 주제를 통한 개념 및 용어로의 하향식 접근과, 구체적인 용어를 통한 개념에의 상향식 접근을 함께 제공하는 체계라고 할 수 있다. 통합 개념체계에서는 이런 점을 고려하여 시소러스를 분류표와 용어사전의 중간에 두고 통합하여 주제와 용어를 개념을 통해 연결시켰다. 이를 통해 주제, 개념, 용어에 대한 상·하향식 접근이 원활하게 이루어지도록 하였다.

통합 개념체계의 각 요소를 통합 연결한 구조도를 간략하게 제시하면 그림 2와 같다.



<그림 2> 통합 개념체계 연결 구조

그림 2에서 점선으로 표시한 부분이 통합 개념체계에 해당하는 부분이다. 표준 분류표의 분류기호가 나타내는 주제에 해당하는 개념이나 용어를 시소러스와 용어사전에서 연계하여 추출할 수가 있으며, 용어사전에 수록된 특정 용어가 나타내는 개념 및 관련 주제를 시소러스와 분류표에서 파악할 수가 있다. 이때 용어 불일치의 문제는 접근 어휘를 통해 해결한다 .

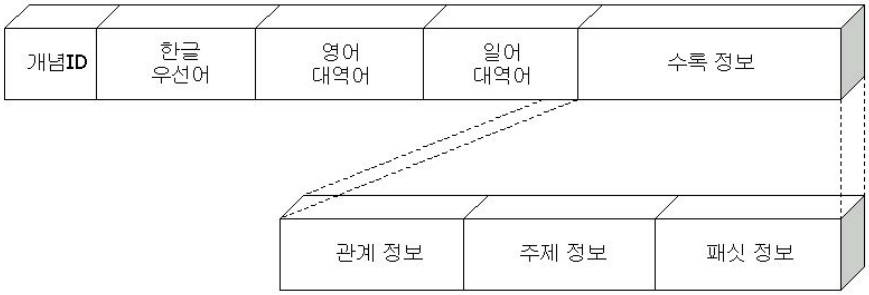
외부 자원인 기존 분류표의 경우에는 표준 분류표와의 매핑 테이블을 통해서 시소러스의 각 개념에 접근할 수가 있다. 기존의 시소러스나 용어사전은 본 연구에서 제시한 형식에 맞추어 개념과 용어에 ID 를 부여하는 방식으로 수록 정보를 통합 개념체계의 일부분으로 포함시킬 수가 있다. 즉, 통합 개념체계가 시소러스, 분류표, 용어사전 등 세 가지 다른 유형의 검색도구를 통합할 뿐만 아니라 복수의 시소러스와 복수의 분류표를 수용할 수 있도록 고려한 것이다.

3. 다국어 시소러스의 설계

통합 개념체계에서 시소러스는 개념을 표현하는 용어를 지시하고 개념간의 관계 정보, 개념의 주제 정보, 개념의 패싯(속성) 정보를 저장한다. 응용 시스템에서 필요로 하는 경우에는 용어사전에서 여러 정보를 추가적으로 도입하여 활용할 수 있다. 특히 용어사전에 수록된 정의 정보의 경우에는 용어의 분명한 의미를 직접적으로 밝힐 필요성을 강조하는 최근 추세(김태수 2001)를 고려해 볼 때 시소러스에서 활용할 가치가 높을 것이다.

3.1. 시소러스의 기본 엔트리

시소러스의 기본 엔트리는 개념을 단위로 한다. 각 엔트리에 대해서 개념 ID를 부여하고, 해당 개념을 대표하는 한국어 우선어(용어사전에서는 표제어)를 지정한다. 영어 및 일본어의 대역어는 각각 영어와 일본어의 우선어로 채택하고, 동일 개념을 뜻하는 다른 용어(비우선어)는 접근 어휘(Access vocabulary)에 수록하여 해당 개념으로의 접근 수단을 제공한다. 시소러스 기본 엔트리의 개괄적 구성은 그림 3 과 같다.



<그림 3> 시소러스 기본 엔트리의 개괄적 구성

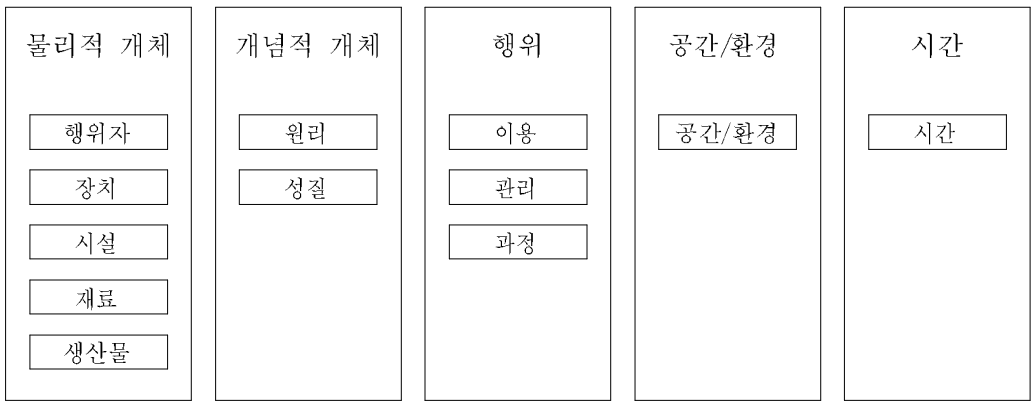
개념ID는 각 개념에 대해서 고유한 식별기호로서 같은 개념을 나타내는 한국어, 영어, 일본어 용어는 모두 동일한 식별기호를 가진다. 동일한 개념을 나타내는 비우선어도 우선어와 같은 개념ID를 가진다.

시소러스의 수록 정보는 관계 정보와 주제 정보, 패싯(속성) 정보의 세 가지로 구성된다.

관계 정보는 계층관계(BT/NT), 연관관계(RT/RT)를 포함한다. 흔히 기존의 시소러스에서 동등관계(USE/UF)로 간주하는 것은 개념간의 관계가 아니라 용어간의 관계이므로 개념 엔트리에 직접 나타내지 않고 접근 어휘에서 처리한다. 일부 사례에서처럼 관계 유형을 세분하는 경우는 표현력은 강화되지만 실제 세분된 관계를 구분하여 지정하기가 쉽지 않으므로 채택하지 않았다. 그대신 패싯 정보를 활용해서 단순한 관계 유형을 보완한다.

주제 정보는 해당 개념의 소속 주제분야를 지시하는 역할을 한다. 소속 주제분야는 기본적으로 본 연구에서 개발한 표준 분류표에 구분된 분류기호로 지정한다. 다른 분류표를 활용할 경우에는 1차적으로는 표준분류표와의 매핑 테이블을 통해서 연결할 수가 있다. 다른 분류표의 분류기호를 더욱 정확하고 세밀하게 지정할 필요가 있는 경우에는 용어사전에서 별도로 처리한다. 예를 들어 '유도가속기'로 표현되는 개념은 표준분류표에서는 분류기호가 D5113(입자가속기)에 해당하며 KDC에서는 429.2, DDC에서는 599.73, LCC에서는 TK 9340에 해당한다.

패싯 정보는 해당 개념의 속성을 나타낸다. 패싯유형은 패싯테이블에 정의된 유형을 사용한다. 패싯유형은 기존 연구를 검토하고 실제 용어들을 살펴본 결과 그림 4와 같이 물리적 개체, 개념적 개체, 행위, 공간/환경, 시간의 다섯 유형을 최상위 패싯으로 설정하고 이를 다시 12개의 유형으로 세분하였다.



<그림 4> 통합 개념체계의 패싯유형

패싯에 대한 기존 연구에서는 '행위자'를 '개체'에 속하게 하지 않고 '이용'이나 '행위'와 함께 모아주는 경우도 있다. 그렇게 할 경우에는 어떤 행위와 그 행위의 주체를 함께 모을 수 있다는 장점은 있으나, 주제 정보와 별도로 본질적인 면에서 패싯 정보를 고려할 경우에는 '행위자'를 '개체'에 속하게 하는 것이 합리적이다.

실제로 시소러스를 구축하면서 패싯 유형을 할당하기 위해서는 해당 주제 분야의 개념들에 대한 각 패싯 유형의 적용 범위와 기준을 미리 검토하여 일관성을 유지해야 한다.

각 개념에 패싯 정보를 할당함으로써 통합 개념체계에서 얻는 이점은 다음과 같다.

첫째, 색인 작업에서 일종의 한정어 역할을 할 수가 있다. 각 용어에 대해서 그 속성을 밝힘으로써 동형이의어 구분에 도움이 된다. 예를 들어 '감자'라는 용어는 주식 분야에서 쓰일 때에는 '관리' 패싯을 가지지만 식물을 뜻할 때에는 '재료' 패싯을 가지므로 구분이 가능하다.

둘째, 검색 작업에서 탐색어들의 조합 관계를 추정하여 검색 성능을 향상시킬 수가 있다. 이용자가 불논리 연산기호를 사용하지 않고 용어 A와 B를 탐색어로 사용하였을 때, A와 B가 동일하게 '재료' 패싯에 해당할 경우에는 'A OR B'에 가깝게 처리하고 하나는 '재료' 패싯, 다른 하나는 '행위' 패싯에 해당할 경우에는 'A AND B'에 가깝게 처리할 수가 있을 것이다.

셋째, 자연어 처리 및 자동 번역 시스템에서 문장을 분석할 때 단어의 패싯을 이용하여 더 정확하게 처리할 수가 있다.

넷째, 연관관계 유형을 다양화하는 효과가 있다. 예를 들어 '이용' 패싯을 가진 용어 A의 관련어 B가 같은 '이용' 패싯을 가지면 유사한 행위 관계로 파악할 수가 있고, 관련어 B가 '재료' 패싯에 해당하면 재료로 이용하는 관계로 파악할 수가 있다.

3.2. 접근 어휘

접근 어휘는 시소러스에 수록된 각 개념에 대한 접근 수단의 역할을 한다. 시소러스에서 각 개념은 그것을 대표하는 각 언어별 우선어가 지정되어 있다. 접근 어휘에는 우선어가 아닌 나머지 한국어, 영어, 일본어의 비우선어를 포함하여 해당 개념 엔트리(우선어)로 지시한다. 다만, 책자형 시소러스에서 한 언어를 기준으로만 자모순 시소러스를 인쇄하였을 경우에는 다른 언어의 우선어도 접근 수단이 있어야 하므로 비우선어와 함께 접근 어휘로 출력하게 된다.

접근 어휘의 엔트리는 접근점과 엔트리 용어로 구성된다. 예를 들어 용어 '오토라디오그래피'는 영어와 일어 대역어로 각각 'autoradiography', 'オートラジオグラフィー'가 해당되지만, 영어 비우선어인 'radioautography'는 시소러스의 기본 엔트리를 구성하지 않고 접근 어휘에 접근점으로 수록되어 한국어, 영어, 일어의 개념 엔트리 용어로 연결되도록 처리한다.

4. 다국어 용어사전의 설계

4.1. 용어사전 모형 비교

Sager(1990)는 용어 데이터베이스의 개념적인 모형을 제안하면서 기본적으로 포함해야 하는 데이터 범주로 표제어(entry term) 및 표제어에 대한 개념적 명세사항, 언어적 명세사항, 화용론적 명세사항을 제시하고 이외에 표제어 정보에 대한 출처 참조 명세사항과 데이터 유지를 위한 관리정보가 필요하다고 하였다.

실제 용어 데이터베이스의 교환을 위한 데이터 요소를 제안하고 있는 TEI P4의 E-TIF와 MARTIF를 비

교해보면 MARTIF가 E-TIF보다 데이터 요소의 구조화 수준이 더 높다는 것을 알 수가 있다. 예를 들어 MARTIF는 용어 데이터베이스에 따라 기술의 수준이 다양한 <term>, <termNote>, <descrip>, <admin> 등의 데이터 요소를 <*Grp> 내에서 묶어주고 있어, 데이터 요소 기술에서의 유동성, 개방성과 구조화 측면을 동시에 보여주고 있다. 또한 용어 데이터 기술을 위해 참조한 정보에 대한 기술의 수준이 TEI P4에 비해 MARTIF가 더 구체적으로 구조화되어 있다.

그러나 이러한 구조화 수준의 차이를 제외하고 본질적인 데이터 요소의 차이는 크지 않다고 할 수 있다. 또한 본질적인 4개의 데이터 요소에 대한 확장 및 한정에 이용되는 속성의 참조는 두 형식 모두 용어 데이터베이스를 위한 데이터 범주를 위한 표준인 ISO 12620을 참조하고 있다. 표1에서 세 가지 레코드 모형을 중심으로 공통된 데이터 요소를 비교하였다.

<표 1> 용어 데이터베이스 모형의 데이터 요소 비교

Sager의 용어 레코드 모형	E-TIF 데이터 요소	MARTIF 데이터 요소
용어(외국어 대응어 포함)	<term>	<term>
용어에 대한 동의어, 약어, 이형	<otherForm>	<termNote>
정의, 맥락	<descrip>	<descrip>
용례 및 용례주기	<descrip>	<termNote>
범위주기	<descrip>	<descrip>
주제분야	※특정 주제코드(DDC 등)를 참조하는 경우 <ptr>로 표시할 수 있음	<descrip>
문법 정보(외국어 대응어 포함)	<gram>	<termNote>
관계	<ref>, <ptr>, <xref>, <xptr>	<descrip>
관리정보	<admin>, <ref>	<admin>

표 1에서 보는 바와 같이 MARTIF의 25개 데이터 요소 중에서 실제적으로 데이터 요소로서의 기능을 하는 것은 <term>, <termNote>, <descrip>, <admin>의 4개 데이터 요소이며, E-TIF의 경우에도 20개 데이터 요소 중에서 9개만이 실제적인 레코드 기입을 위해 이용되는 데이터 요소라는 것을 알 수 있다. 각각에서 중복되는 데이터 요소들은 각 항목을 상세하게 기술하기 위해 ISO 12620의 데이터 범주를 이용한 'type' 속성을 이용하여 데이터의 내용을 자유롭게 한정 및 확장할 수 있다.

실제적인 데이터 기입을 위해 필요한 데이터 요소 이외의 나머지 요소들은 기본 데이터 요소의 기술을 위해 부가적으로 이용되는 문서 구조의 표현이나 데이터 기술 시 필요한 정보를 담기 위해 이용된다. 특히 MARTIF의 <item>이나 <refObject>, <ptr>, <ref> 류, E-TIF의 <ref>, <ptr> 류와 같은 데이터 요소는 용어 엔트리에 기입된 내용을 참조한 정보를 기술하기 위해 이용되는 데이터 요소이다.

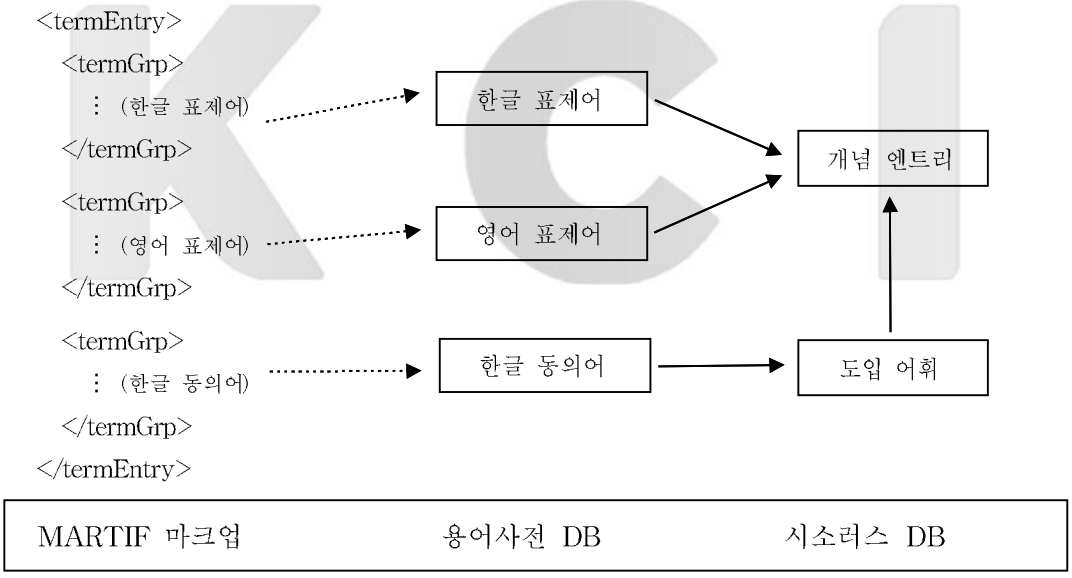
본 연구의 용어사전 데이터베이스의 형식은 이상 살펴본 세 가지 용어 데이터베이스 형식을 참고로 하되 가장 상세하고 구조화가 잘 된 MARTIF 표준을 주된 근거로 설계하였다. MARTIF 표준을 근거로 할 때에는 데이터 속성을 한정 및 확장하기 위해 참조한 데이터 범주 표준인 ISO 12620을 동일하게 참조하되, 전체를 참조하는 것이 아니라 필수요소와 임의요소로 나누어 구성하기로 한다.

4.2. ISO 12620 표준의 구성

용어사전 데이터베이스의 데이터 요소를 설계하기 위해 MARTIF와 TEI P4에서 전문용어 데이터 범주의 정규화를 위한 표준으로 삼고 있는 ISO 12620(Computer Applications in Terminology - Data Categories)을 참조하였다.

다양한 전문용어 데이터베이스 환경에서는 데이터 범주를 표현하기 위해 상이한 필드명을 이용하기 때문에 데이터 교환을 위해서는 정규화된 형식으로 이들을 대체할 필요가 있다. ISO 12620은 기계가독형 전문용어 교환을 위한 용어 데이터의 표준 범주를 제시하고 있는 표준으로, 다양한 용어학적 데이터 항목을 상세하게 나타내고 있다.

또한 ISO 12620은 MARTIF 및 다른 용어 교환 형식에서 추구하는 개념단위 기술을 지원한다. 개념단위 기술을 원칙으로 하는 MARTIF에서 하나의 개념을 설명하기 위해 모든 용어를 복수의 <termGrp> 내에서 표시하고, 그 전체의 개념은 <termEntry> 내에 포함되도록 구성하고 있는데, 이러한 개념단위 기술을 표현하기 위해 ISO 12620에서는 용어 범주에서 하나의 개념과 관련된 모든 용어를 표시하고, 용어 관련 범주에서 하나의 개념을 이루는 여러 용어의 특성을 설명하도록 되어 있다. MARTIF를 이용한 개념단위 기술을 용어사전 DB와 시소러스 DB의 개념 표현 체계에 대응시킨 것을 그림 5에 나타내었다.



<그림 5> MARTIF와 통합 개념체계 DB의 개념 단위 기술 대비

ISO 12620은 10개의 데이터 범주를 표 2와 같이 크게 3개의 범주로 나누어 설명하고 있다.

<표 2> ISO 12620 표준의 기본 구조

범주	하위 범주
용어 및 용어 관련 데이터 범주	- 용어: 전문용어 데이터베이스에 출현하는 용어 - 용어관련정보: 용어의 특정한 본질이나 특성을 기술 - 번역대응어: 동일하거나 매우 유사한 개념에 부여된 다른 언어로 된 용어간의 번역대용어와 관련된 데이터
개념 기술 관련 데이터 범주	- 주제 분야: 용어에 대한 분류 정보 - 개념 관련 기술: 정의, 설명, 문맥 등의 개념 기술을 위한 데이터 - 개념 관계: 특정 용어에 대한 개념쌍 간의 개념관계를 설명 - 개념 구조: 특정 용어에 대한 개념 구조를 설명 - 주기: 개념 기술과 관련된 별도의 주기 정보
관리 데이터 범주	- 문서 언어: 용어 데이터베이스 작성과 관련된 시소러스 정보 - 관리 정보: 용어 데이터베이스 관리 정보

4.3. 용어사전 구성

통합 개념체계 안에서의 용어사전 엔트리는 ISO 12620 표준을 고려하여 용어 및 용어 관련 데이터, 개념 기술 관련 데이터, 관리 데이터의 세 부분으로 구성하였다. 구체적으로 ISO 12620 표준에서 각 부분에 포함하는 다양한 데이터 요소 중에서 그림 6과 같이 19개의 필수요소를 선정하여 용어사전을 위한 표준 요소를 구성하였다. 필수요소에는 포함되지 않았지만 ISO 12620 표준에 정의된 다른 요소들은 각 기관이 용어사전을 구축하거나 교환할 때에 자관의 특성에 맞게 ISO 12620 표준을 참고로 하여 확장이 가능한 임의요소로 하였다.

용어 및 용어 관련 데이터						
용어	용어 유형	완전형	생략형	이형	품사	문법적 수

개념 기술 관련 데이터		
분류체계	분류번호	정의

관리 데이터								
시소러스명	입력일	수정일	입력자	갱신자	언어코드	개념 식별기호	용어 식별기호	정의 출처 식별기호

<그림 6> 용어사전 엔트리의 필수요소

필수요소 각각에 대한 설명은 다음과 같다.

1) 용어 : 특정 언어로 표현되는 단일 개념에 대한 명시로, 단일 개념을 표현하기 위해 용어사전 내에 수록된 모든 용어(특정 용어에 대한 완전형, 생략형, 동의어, 이형 등)와 그에 대응하는 번역대응어 등을 포함한다. 용어는 단일어나 복합어의 형태를 지닌다. 한국어의 경우 복합명사로 구성된 표제어는 붙여쓰기를 원칙으로 하고 띄어써서 복합어가 될 경우의 형태는 이형표기에서 다룬다.
2) 용어 유형: 용어 레코드 내에서 특정 개념을 설명하는 여러 용어 중 표제어인지 동의어인지 여부를 밝혀준

다. 시소러스에서 우선어, 비우선어 정보에 해당하는 항목이다. 원래 교환용 표준형식인 ISO 12620에서는 해당 용어가 표제어인지 여부, 혹은 동의어인지 여부를 별도의 속성으로 각각 나타내지만 여기서는 하나의 필드에서 표제어는 'M', 동의어는 'S'로 나타내도록 하였다. 통합 개념체계는 다국어 엔트리를 갖는 용어사전이므로 각 언어별로 표제어가 있다. 예를 들어 개념ID C000000185는 한국어로는 '니오븀'이 표제어, '니오브'가 동의어이며, 영어로는 'niobium'이 표제어, 'columbium'이 동의어이고, 일본어로는 'ニオブ'이 표제어이다. 한국어 표제어의 선정은 시소러스의 우선어 선정기준에 따랐으며 다른 언어의 표제어 선정은 한국어 표제어와 의미가 가장 가까운 대역어로 지정하는 것을 원칙으로 하였다.

3) 완전형 : 특정 용어의 성격이 생략형인 경우에 해당 용어의 완전형을 표기한다.
4) 생략형 : 특정 용어의 성격이 완전형인 경우에 해당 용어의 생략형을 표기한다.
5) 이형 : 용어의 유형을 지시하기 위한 것의 일종으로, 용어 엔트리에 대한 특정한 대체 형식이 있음을 표현하기 위해 이용된다. 이형의 종류에는 철자 이형과 띄어쓰기 이형이 있다. 한국어는 전문 용어의 경우에 복합명사의 형태로 붙여쓰거나 복합어의 형태로 띄어쓰는 것이 둘 다 허용되므로 용어의 기본형은 붙여쓴 형태로 하고 띄어쓴 이형을 여기에서 표기한다. 예를 들어 용어 '수은증기냉각로'의 경우에는 '수은 증기 냉각로'와 같은 형태를 적어준다.
6) 품사 : 용어의 문법적인 속성에 기초하여 용어에 대한 품사 정보를 기술한다. 품사에 대한 하부 속성으로는 명사(n),동사(v), 형용사(a), 기타(o)의 4가지를 포함한다.
7) 문법적 수 : 용어가 표현하고 있는 객체의 수를 지적하기 위해 문법적인 지시를 따르는 것으로, 그 하부 속성으로는 단수(s)와 복수(p)를 포함한다.
8) 분류체계 : 용어 데이터베이스 내의 개별 용어에 대한 주제정보나 분류번호를 제공하기 위해 참조한 분류체계를 명시한다. 용어사전에 명시된 분류체계는 통합개념체계 내에서 분류표로 연결되는 포인터가 된다. 예를 들어 용어 데이터베이스 내의 개별 용어에 대한 분류번호를 부여하기 위해 DDC를 이용한 경우, 분류체계는 'DDC'가 된다.
9) 분류번호 : 용어에 대한 주제적인 표현을 위해 사용하고 있는 특정 분류체계 내에서의 해당 용어의 주제적인 위치를 표현하기 위해 이용한다. 예를 들어 용어 '압력관형원자로'에 대해서 DDC 분류체계를 이용하여 주제적인 위치를 표현하면 '621.4834', LCC 분류체계를 이용할 경우에는 'TK 9202'가 된다.
10) 정의 : 다른 개념과의 차이를 보여주거나 특정 개념에 대한 설명을 위한 기술을 의미한다. 특정 개념에 대한 정의는 해당 주제분야에서 권위적인 2차 정보원을 참조하거나, 주제적으로 잘 알고 있는 용어 레코드 입력자가 직접 해당 개념에 대한 정의를 입력하는 방법을 이용할 수 있다. 2차 정보원을 참조한 경우에는 필수 요소인 '출처'에 그에 대한 서지 사항을 입력해야 한다.
11) 시소러스명 : 용어사전 구성을 위해 용어 엔트리나 디스크립터의 출처로 사용한 시소러스의 명칭을 밝힌다.
12) 입력일 : 용어 레코드를 입력한 날짜를 기록한다. YYYY-MM-DD의 형식을 따른다.
13) 수정일 : 용어 레코드를 수정한 날짜를 기록한다. YYYY-MM-DD의 형식을 따른다.
14) 입력자 : 용어 레코드의 질적 책임을 기록하기 위해 입력자 정보를 기록한다. 입력자의 이름이나, 전화번호, 전자우편 등의 연락처를 입력하거나, 이러한 정보를 코드화하여 기록할 수 있다.
15) 갱신자 : 용어 레코드의 질적인 책임을 기록하기 위해 갱신자 정보를 기록한다. 갱신자의 이름이나, 전화번호, 전자우편 등의 연락처를 입력하거나, 이러한 정보를 코드화하여 기록할 수 있다.

16) 언어 코드 : 용어 레코드의 용어 표현에 사용된 언어에 대해 ISO 639에서 지정한 두 자리 표준 코드를 이용하여 기록한다.
17) 개념 식별기호 : 개념 단위로 용어 데이터 레코드를 식별하기 위한 코드이다. 데이터베이스에 저장될 때 자동으로 생성할 수 있다. 동일 개념을 나타내는 모든 용어를 연결해주는 역할을 한다. 통합 개념체계 전체적으로는 개념 식별기호를 통해 용어사전에서 시소러스로 연결된다.
18) 용어 식별기호 : 용어사전 내 각 용어에 대한 고유 식별기호를 부여하기 위한 코드이다. 데이터베이스에 저장될 때 자동으로 생성할 수 있다.
19) 정의 출처 식별기호 : 용어의 정의 작성을 위해 참조한 정보원에 대한 서지사항을 식별하기 위한 코드이다. 이 코드를 이용하여 서지정보의 완전형과 연결함으로써 출처에 대한 완전 서지정보를 얻는다. 실제 참고정보원의 서지사항은 별도의 참고정보원 테이블로 관리한다.

위와 같은 표준 데이터 항목을 기준으로 '열중성자화'라는 개념에 대하여 실제 데이터를 입력한 예는 그림 7과 같다.

← 용어 및 용어 관련 데이터 →							← 개념 기술 관련 데이터 →		
열중성자화	M		열화	열 중성자화	n	s	ICC	D5110	고속중성자가 감속과정을 거쳐 매질과 열평형에 도달하여 열중성자가 되는 것.
ICT	2001.11.20		정영미		kr	C000000698	TK00000698	KND1996	
← 관리 데이터 →									
열화	S	열중성자화			n	s	ICC	D5110	고속중성자가 감속과정을 거쳐 매질과 열평형에 도달하여 열중성자가 되는 것.
ICT	2001.11.20		정영미		kr	C000000698	TK00001578	KND1996	
thermalization	M				n	s	ICC	D5110	고속중성자가 감속과정을 거쳐 매질과 열평형에 도달하여 열중성자가 되는 것.
ICT	2001.11.20		정영미		en	C000000698	TE00000698	KND1996	
熱中性子化	M				n	s	ICC	D5110	고속중성자가 감속과정을 거쳐 매질과 열평형에 도달하여 열중성자가 되는 것.
ICT	2001.11.20		정영미		jp	C000000698	TJ00000698	KND1996	

<그림 7> 용어사전 레코드의 예

5. 과학기술 분류표의 구성

과학기술분야에서 사용되는 일반분류표에는 한국십진분류법(KDC), 듀이십진분류표(DDC), 미국의회도서관분류표(LCC), 국제십진분류표(UDC) 등이 있으며, 전문분류표에는 국제특허분류표, NLMC(미국 National Library of Medical Classification), JICST 분류표 등이 있고, 우리나라에는 과학기술문헌 DB 분류표(산업기술정보원), KISEP 분류표, 해외과학기술동향 DB 분류표(KORDIC), 건설정보분류체계(한국건설기술연구원), 기술정보분류집(현대건설), 과학기술 분야분류(한국과학재단) 등이 있다. 현재 우리나라의 과학기술계에서는 각 분야별로 학계 및 현장의 전문가들을 통하여 지속적인 분류표 개발을 도모하고 있다.

개념 통합체계에 대한 주제 접근을 위해서 본 연구에서는 기존 분류표들을 검토한 결과 가장 최근에 개발된 KISEP 분류표에 기반하여 종합적인 과학기술 전문 분류표(ICC: Integrated Concept Classification)를 제안하였다.

신규 분류체계(ICC)의 작성 원칙은 다음과 같다.

- 백진법을 적용하여 주류와 하위류들을 각각 99개 항목으로 구분할 수 있도록 한다.
- 가능하면 기호의 길이를 간단하게 한다.
- 기호구성을 단순하게 한다.
- 도서관 및 DB관리에서 사용하기 용이하게 한다.
- 기호에 신축성이 있게 하여 신규 주제의 첨가를 가능하게 한다.
- 조기성을 유지하도록 한다.
- 전문분류표로서 과학기술 주제가 상세하게 전개될 수 있도록 작성한다.
- 주제 배열은 학문의 발전과정을 참고하여 기초부터 최신 기술 순서로 작성한다.
- 새로이 출현한 주제 및 개념들을 위한 분류항목을 신설할 수 있도록 유의한다.
- 기호는 영역분류, 대·중·소 분류로 한정하고 세목전개는 후의 과제로 남긴다.
- 추후에 과학기술 외의 주제를 첨가하여 종합적 분류표로서도 기능할 수 있게 한다.

기본기호 구성의 기본단위는 7단위로 하되 최상위 영역분류는 알파벳 대문자 1 개를 사용하여 표 3 과 같이 정한다. 그 아래의 여섯 자리 숫자는 두 자리씩 묶어서 대분류, 중분류, 소분류 수준의 세 단계 백진법 분류를 전개하였다.

<표 3> 최상위 영역분류

기호	항 목	범 위
A	자연과학 (Natural Sciences)	자연과학, 순수과학 분야
B	생명과학 (Life Sciences)	생물학, 식품과학, 농학, 축산학, 수산학 등
C	의과학 (Medicial Sciences)	의학, 간호학, 약학 등
D	기술, 공학 (Technology, Engineering)	기술, 공학 분야
M	제조업 (Manufactures)	제조업 분야
Z	과학기술의 연계학 (Interdisciplinary Sciences)	과학, 기술과 일반적으로 관련된 분야

예를 들어서 '중성자 빔기술'에 대한 주제 계층은 다음과 같이 구성된다.

D 공학 ····· 제 1 단위 (영역분류)
D51 원자력공학 ····· 제 2,3단위 (대분류)
D5112 연구용 원자로 ····· 제 4,5단위 (중분류)
D511201 중성자 빔기술 ····· 제 6,7단위 (소분류)

이와 같이 구성한 통합 개념체계를 위한 과학기술분야 분류표에 대해서 각 분류항목에 해당되는 KDC, DDC, LCC 번호를 파악하여 연결하는 매핑 테이블을 작성하였다.

주제 분류에 더하여 형태 정보와 지역 정보를 추가할 필요가 있는 경우에는 기본 번호에 형태구분 기호와 지역구분 기호를 추가할 수 있게 하였다. 자료 형태에 의한 구분은 알파벳 소문자를 첨가하여 표현하게 하고, 자료의 지역적인 특성을 나타낼 때에는 기본번호에 g(geographical treatment)를 첨가한 후에 KDC의 지역구분에 의거한 지역구분기호를 첨가하도록 하였다.

6. 원자력분야 프로토타입 구축

6.1. 시소러스 구축

원자력분야를 대상으로 통합 개념체계 프로토타입을 구축하기 위해서 원자력 분야의 핵심 개념을 나타내는 용어를 수집하여 시소러스를 구축하였다. 구축된 시소러스에는 원자력분야 세부 분류표와 패싯유형을 참조하여 주제정보와 패싯유형 정보를 수록하였다.

원자력분야의 대표적인 시소러스인 INIS Thesaurus(International Atomic Energy Agency 1999)와 국내의 『최신 원자력용어사전』(한국원자력산업회의 1996)을 용어 수집에 이용하였다. 1999년판을 기준으로 INIS Thesaurus의 우선어는 19,007개, 비우선어는 6,676개이다. 여기에는 일반물리학, 고에너지물리학, 핵물리학, 플라즈마물리학, 원자 및 거대분자물리학, 화학, 물질, 지구과학, 방사선생물학, 방사성동위원소 효과, 동력학, 응용생물학, 방사선 보호 및 환경, 방사선학, 원자력의학, 동위원소, 방사선원, 사회학, 원자력법, 안전 및 검사, 비원자력 경제 및 환경효과 등 폭넓은 주제분야에서 관련 용어가 선정되어 있다. 또한 원자력과 밀접한 관련이 없는 일반 용어도 상당수 선정되어 있다.

한편 『최신 원자력용어사전』의 수록어휘는 5,500개(1996년판 기준)로서 원자력과 관련된 기초 및 응용과학, 기술, 제도, 환경문제 등에 관련된 용어가 선정되어 있다. 두 자료의 일치율을 하부 주제 두 가지를 대상으로 검토해본 결과 미일치되는 경우가 매우 많았으며, 전체적인 일치율은 시소러스 기준으로는 약 5-6%, 용어사전 기준으로는 약 20-25%로 나타났다.

실제 용어의 선정은 다음과 같이 하였다. 우선 두 자료 사이에 일치된 용어(약 1,150개)를 우선 선정하였다. 그런 다음 INIS Thesaurus에만 있는 용어 중 고빈도 출현용어(원자력문헌에 자주 출현하는 용어) 346개를 추가 선정하였다. 선정된 용어들간의 관계를 검토하여 계층관계가 없거나 끊어진 경우의 연결을 위해, INIS Thesaurus상의 NT 및 RT 용어를 대상으로 출현빈도수를 고려하여 추가로 용어를 92개 선정하였다. 이중에서 다시 원자력분야와 밀접하지 않은 개념, 원소 및 너무 일반적인 개념의 용어 103개를 원자력분야 전문가와 함께 선정한 후, 출현빈도 및 계층관계 등을 고려하여 최종적으로 63개의 용어를 제외하였다. 그 결과 선정된 최종 용어는 1,526개가 되었다.

우선어의 선정시에는 일단 INIS Thesaurus를 참고로 하여 영어를 기준으로 한 다음, 이 용어가 나타내는 개념에 해당하는 한국어 용어와 일본어 용어를 결정하고 각각을 한국어, 영어, 일어 우선어로 지정하였다.

원자력 분야에서 패싯 유형을 적용할 때 각 패싯에 대해서 고려한 할당 대상의 범위와 실제 할당한 사례는 표 4와 같다.

<표 4> 원자력분야 용어에 대한 패싯의 적용

패싯 유형		적용 범위	대상 용어 사례	할당된 개념 수
물리적 개체	행위자	개인, 조직, 기구, 생물, 생체기관	원자로운전사, IAEA, 생식세포	10
	장치	기기, 도구, 구조, 노심	감마카메라, 고속로, 방호복	249
	시설	플랜트, 발전소, 공장	저장시설, 원자력발전소, 중수공장	16
	재료	원소, 핵연료, 요소재료 ,전자기파	농축우라늄, 강, 감마선	396
	생산물	산출물, 결과물, 가공물	방사성폐기물, 오염물질, 정광	14
개념적 개체	원리	원리, 법칙, 이론, 모형	핵물리학, 질량공식, 군상수	68
	성질	성질, 상태, 조건, 속성, 열, 힘, 양	수명, 질량, 초전도, 휨성	258
행위	이용	사용, 측정, 조사, 치료, 생산	온도감시, 음향방출시험, 일시피폭	78
	관리	평가, 안전관리, 폐기물 및 핵연료 관리, 제도 및 정책, 법률	사찰, 안전규제, 부지선정, 폐기물처분	48
	과정	현상, 상태 변화(또는 상태변화를 일으키는 방법), 경로, 질환	감속, 광자방출, 방사, 약한상호작용	353
공간/환경	공간/환경	주변상황, 위치	방사선관리구역, 생태계, 환경	5
시간	시간			0

원자력분야 분류표는 세분류(4단계)까지 전개하였으나, 용어 개념의 포괄성 및 망라성으로 인해 소분류(3단계) 수준의 번호를 각 용어에 부여하였다. 하나의 용어에 하나의 분류번호를 부여하는 것을 원칙으로 하였으나 불가피한 경우에 한해서는 2개까지를 허용하였다. 원자력과 관련한 핵물리분야 및 원소는 '원자력기초기술'에 분류하고, 적절한 소분류번호를 부여할 수 없는 용어는 해당 중분류의 '미분류' 및 '기타'에 분류하였다.

6.2. 용어사전 및 세부 분류표 구축

원자력 분야 용어사전의 각 용어에 대한 정보를 수집하기 위해서 원자력과 화학 분야, 그리고 과학기술 일반 분야의 국내외 사전을 정보원으로 하였다. 용어사전에는 정의 정보의 출처를 밝혔는데, 각 출처에 대한 완전한 서지사항 대신 기호를 표기하였다. 용어 정의를 기존 사전에서 찾을 수가 없는 용어는 원자력 분야 전문가에게 문의하여 정의를 수록하였다. 확실한 정의 문장을 입수하지 못한 일부 용어는 정의 항목을 비워두었다.

정의 정보는 각 사전에 수록된 용어의 설명 중에서 정의항과 피정의항의 개념 범주, 정의항과 피정의항의 관계, 피정의항의 본질적 특성을 나타내는 항목을 선별하여 1 문장에서 5 문장 사이로 수록하는 것을 원칙으로 하였다. 대부분의 용어 정의는 두 세 문장으로 구성되었다.

이형 표기는 한국어의 경우에 복합명사를 복합어의 형태로 띄어쓴 형태를 입력하였다. 이때 모든 형태소를 띄지 않고 가급적 띄어진 단어가 개별 용어로 사용될 경우에만 띄어썼다. 특히 원자력분야에서 한 글자 용어의 경우에는 다음의 8가지만 이형표기에서 분리하였다.

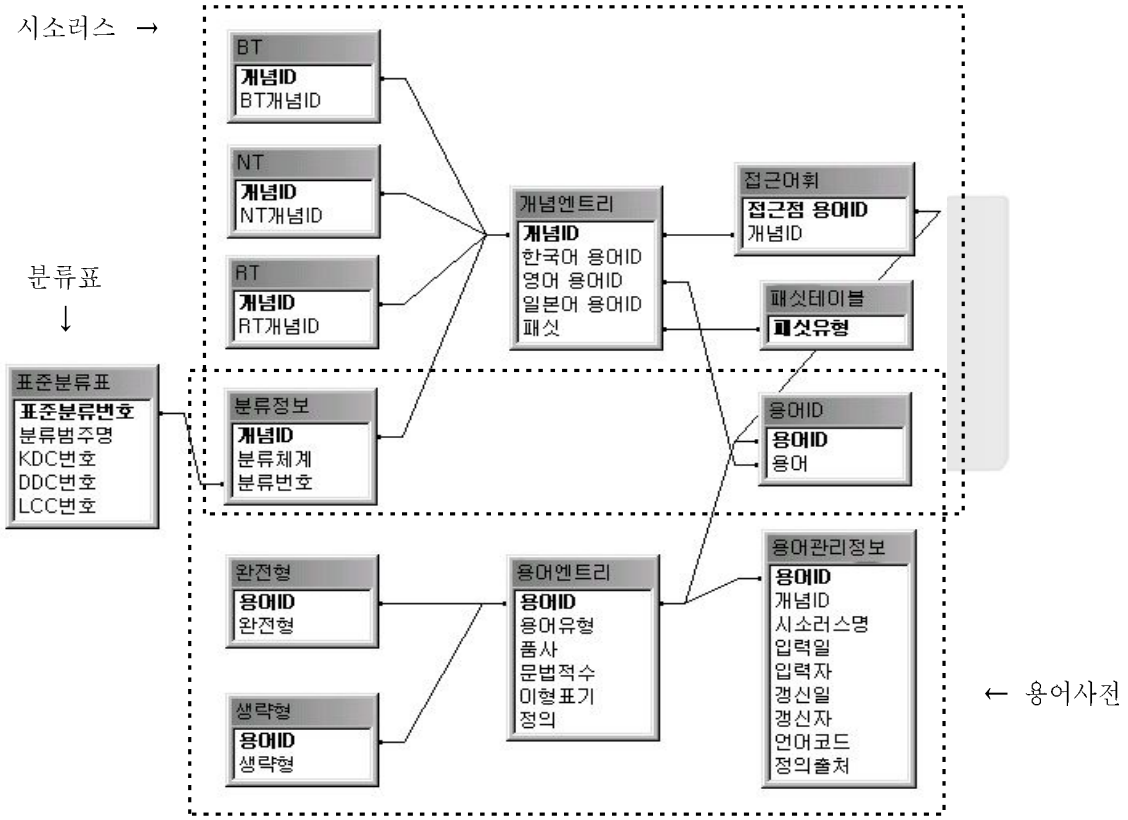
핵, 열(heat), 빔(beam), 총(gun), 점(point), 각(angle), 강(steel), 홀(hall)

원자력 분야는 과학기술 표준분류표(ICC)에서 D51에 해당하는데 세부 주제를 다음과 같이 6 개로 구분하고 이를 다시 소주제로 세분한 다음 각 항목에 대한 KDC, DDC, LCC 번호 매핑 테이블을 작성하였다.

- 원자력 기초기술 (D5110)
- 원자로, 원자로 공학(D5120)
- 핵연료 주기 및 방사성폐기물 관리(D5130)
- 원자력 안전(D5140)
- 방사선 방호 및 이용 (D5150)
- 원자력 시스템엔지니어링(D5160)

6.3. 통합 데이터베이스 구현

앞에서 구축한 시소러스 및 용어사전 데이터를 통합하여 원자력분야의 통합 개념체계 데이터베이스를 그림 8과 같이 구현하였다.



<그림 8> 통합 개념체계 데이터베이스 구성

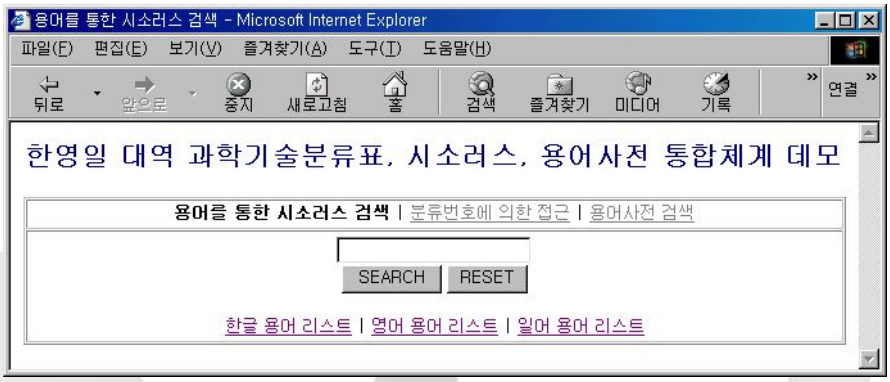
시소러스의 주 테이블은 개념 엔트리로 구성하였고, 용어사전의 주 테이블은 용어 엔트리로 구성하였다. 각 용어에 부여된 용어ID 정보는 용어사전의 주요 정보이긴 하지만 시소러스에서도 기본적으로 참고되므로 양쪽에 동시에 속하게 하였다. 한편 분류정보는 개념을 단위로 할당되므로 시소러스의 주요 정보이지만 용어사전에서도 각 용어에 상응하는 개념 엔트리를 통해서 기본 정보로 참고하므로 역시 양쪽에 속한 것으로 하였다.

데이터베이스 구축에는 MS-Windows 98을 운영체제로 하는 Pentium III PC 플랫폼에서 MS-Access를 도구로 사용하였다. 일본어 용어의 경우에는 윈도 운영체제에서 지원하여 MS

Access에서 사용되는 일본어 코드로 저장하였으며, 윈도 계열이 아닌 계열의 LINUX 운영체제나 월드와이드웹에서 활용할 수 있도록 Shift-JIS 코드 표기를 별도의 보조 테이블에 저장해두었다.

6.4. 통합 데이터베이스 검색

통합 개념체계는 시소러스와 용어사전, 분류표를 연동시켜 활용할 때 그 가치를 십분 발휘하게 된다 구축된 통합 데이터베이스에 수록된 개념과 용어에 대해서 다양한 방식으로 접근하는 사례를 아래에 제시하였다. 검색 시스템은 Pentium III 800Mhz CPU의 LINUX 서버에서 Apache 웹 서버와 mySQL 데이터베이스 서버, PHP 스크립트를 이용해서 구현하였다. 그림 9는 검색 시스템의 첫 화면으로서 용어를 통한 시소러스 검색, 분류번호에 의한 접근, 용어사전 검색의 세 가지 주 메뉴를 제공한다

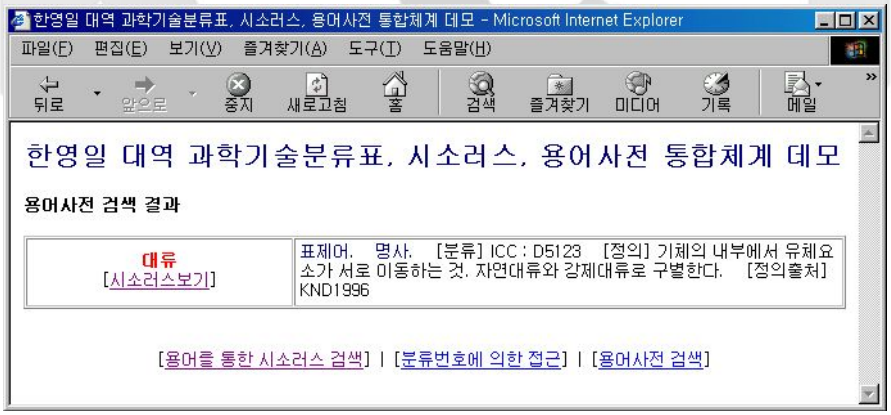


<그림 9> 통합 개념체계 검색 데모 시스템 초기 화면

그림 9의 '용어를 통한 시소러스 검색' 메뉴에서는 한국어, 영어, 일본어 용어를 직접 입력하거나, 또는 화면 하단의 해당 용어리스트를 선택하여 나타나는 용어리스트를 통해서 해당 개념 엔트리에 접근할 수가 있다.

'대류'에 대한 개념 엔트리를 검색한 결과는 그림 10과 같다. 여기서 구체적으로 각 용어에 대해서 살펴보려면 용어 옆의 '사전보기' 링크를 통해서 그림 11과 같은 용어 엔트리로 접근할 수가 있다. 이 개념에 부여된 분류기호 'D5123'을 통해서는 그림 12와 같이 동일한 주제로 범위를 넓혀서 해당하는 개념들을 찾아볼 수가 있다.

<그림 10> 시소러스 개념 엔트리 검색 결과



<그림 11> 용어사전 검색 결과

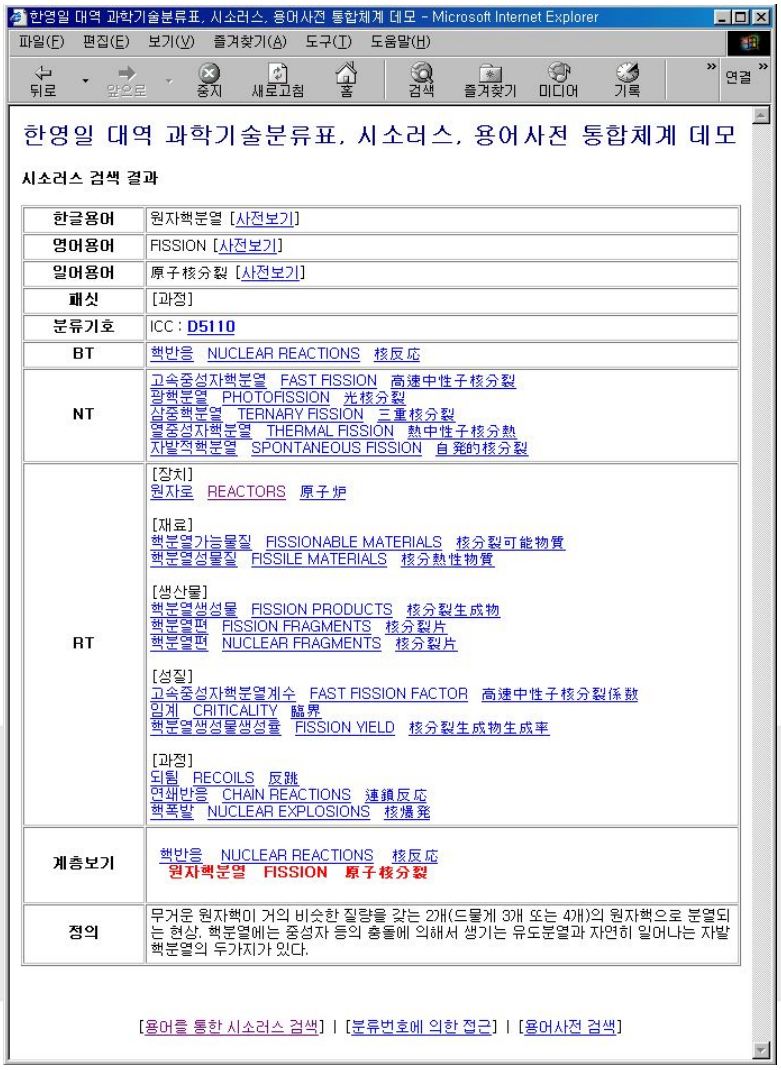
<그림 12> 동일 주제 개념 엔트리 검색 결과

분류표를 통해서 접근할 때에는 그림 13과 같이 분류번호를 직접 입력하거나 분류표를 보고 항목을 선택할 수가 있다. 이때 표준 분류표만 아니라 매핑 테이블이 제공되는 다른 분류표의 분류번호로도 검색하여 해당 주제에 포함된 개념 엔트리 집합에 접근할 수가 있다.

분류번호	하위분류번호	분류범주명	KDC	DDC	LCC
D5100	00	원자력공학 Nuclear Engineering	559.1	621.48	TK 9001-9401
D5101	00	원자력 정책 및 산업 Nuclear Policy and Industry	559.1	354.47	TK 9145
D5101	01	원자력 정책 및 행정 Nuclear Policy and Administration	559.1	354.47	TK 9145
D5101	02	원자력 산업 Nuclear Industry	559.1	338.47	TK 9145
D5110	00	원자력 기초기술 Basic Nuclear Technology	559.1	621.48	TK 9155.5
D5111	00	핵융합 Nuclear Fusion Technology	559.13	621.484	TK 9204
D5111	01	핵융합플라즈마공학 Nuclear Fusion Plasma Technology	559.13	621.484	TK 9204
D5111	02	핵융합 반응장치 Nuclear Fusion, Reactor	559.13	621.484	TK 9204
D5111	03	관성 핵융합 Interia Nuclear Fusion	559.13	621.484	TK 9204
D5111	04	플라즈마 응용기술 Plasma Application Technology	559.13	621.484	TK 9204
D5112	00	연구용 원자로 Research Reactor	559.1245	621.4834	TK 9203.B6
D5112	01	중성자 빔 이용 Neutron Beam Technology	559.1245	621.4834	QC 793.5
D5112	02	핵연료 조사 Nuclear Fuel irradiation Testing	559.125	621.48335	TK 9360
D5112	03	원자력재료 조사 Nuclear Materials Irradiation Testing	559.123	621.4833	TK 9185.A2
D5112	04	방사화 분석 Radioactivation Analysis	431.48	543.0882	TK 9400
D5112	05	중성자 라디오그래피 Neutron Radiography	429.2213	539.7213	QC 793.5
D5113	00	입자가속기 Particle Accelerators & Applications	429.2	539.73	TK 9340
D5113	01	전자 가속장치 Electron Accelerators	429.234	539.73	QC 787.E39
D5113	02	양성자 가속장치 Proton Accelerators	429.234	539.73	QC 787.p7

<그림 13> 분류표 및 분류번호 검색을 통한 접근

시소러스 개념 엔트리의 검색 결과에 나타나는 해당 개념의 연관개념들은 그림 14와 같이 패싯을 기준으로 묶어서 제시하여 연관관계를 구체적으로 세분하는 효과를 낼 수도 있다. 그림 14의 예에서 연관개념을 보면 '원자핵분열'이라는 [과정]은 '원자로'라는 [장치]를 통해서 '핵분열성물질'을 [재료]로 하여 '핵분열생성물'을 [생산물]로 산출하는 것으로서 '핵분열생성물생성률'과 같은 [성질]이 있다는 것을 자연스럽게 기계적으로 유추할 수가 있다.



<그림 14> 패싯을 이용한 연관개념의 구분

6.5. 통합 데이터베이스 출력

데이터베이스가 아닌 책자 형태로 이용할 경우를 위해서 통합 개념체계 데이터베이스의 내용을 인쇄 출력하는 형식을 정하고 원자력 분야 자모순 시소러스, 계층적 시소러스, 접근 어휘, 용어사전을 인쇄하였다.

1) 자모순 시소러스 출력

각 개념 엔트리를 한국어 용어의 자모순 기준으로 출력하였다. 필요에 따라서는 영어나 일본어의 자모순으로 출력할 수도 있을 것이나, 여기서는 영어 및 일본어 우선어를 접근 어휘에 포함시켜서 해당 개념 엔트리로 접근할 수 있도록 하였다. 엔트리 용어의 출력은 3개 국어를 함께 출력하되 한국어 용어를 맨 앞에 표기하였다. 한국어 용어는 띄어쓴 형태를 용어사전에 저장하고 있으므로 복합어라 할 지라도 띄어쓰기 없이 모두 붙여서 제시하였다. '방사성동위원소'라는 개념 엔트리의 출력 예는 그림 15와 같다.

우선어 → 방사성동위원소 RADIOISOTOPES 放射性同位元素
패싯 → FA [재료]
분류정보 → CL ICC : D5152
접근어휘 → UF 방사성동위체
RADIONUCLIDES
상위개념 → BT 동위원소 ISOTOPES 同位元素
하위개념 → NT 골수핵종 BONE SEEKERS 新骨性物質
베타붕괴방사성동위원소 BETA DECAY RADIOISOTOPES β崩壊放射性同位元素
자발적핵분열동위원소 SPONTANEOUS FISSION RADIOISOTOPES 自発的核分裂同位元素
관련개념 → RT 방사능 RADIOACTIVITY 放射能
방사면역분석시험 RADIOIMMUNOASSAY 放射免疫分析試験
방사선원 RADIATION SOURCES 放射線源
방사선핵종이동 RADIONUCLIDE MIGRATION 放射線核種移動
방사성물질 RADIOACTIVE MATERIALS 放射性物質
방사성핵종관리 RADIONUCLIDE ADMINISTRATION 放射性核種管理
방사성핵종동력학 RADIONUCLIDE KINETICS 放射性核種動力学
핵의학 NUCLEAR MEDICINE 核医学

<그림 15> 시소러스 엔트리의 출력 예

2) 계층적 시소러스 출력

계층적 시소러스에서는 시소러스의 엔트리 계층 체계를 출력하였다. 최상위 개념을 기준으로 하위 개념을 한 단계 씩 마침표(.) 기호로 들여쓰기를 하여 그림 16과 같이 구분하였다. 최상위 개념을 비롯하여 각 엔트리의 정렬은 한국어 용어를 기준으로 하였다.

연료 FUELS 燃料
. 화석연료 FOSSIL FUELS 化石燃料
. 액체연료 LIQUID FUELS 液体燃料
. . 핵연료용액 FUEL SOLUTIONS 核燃料溶液
. . 액체금속핵연료 LIQUID METAL FUELS 液体金属核燃料
. 핵연료 NUCLEAR FUELS 核燃料
. . 고체연료 SOLID FUELS 固体燃料

<그림 16> 시소러스 계층의 출력 예

3) 접근 어휘 출력

접근 어휘는 한국어 우선어를 제외한 한국어, 영어, 일본어의 비우선어와 영어, 일본어의 우선어를 자모순으로 출력하고 오른쪽에 해당 개념 엔트리 용어를 제시하였다. 영어와 일본어 용어는 우선어와 비우선어를 함께 자모순으로 출력하되, 우선어는 왼쪽에 '+' 기호를 출력하여 구분하였다. 그림 17 에 '방사성동위원소' 와 관련된 접근 어휘의 출력 예를 보였다.

⋮ ⋮

방사성동위체 방사성동위원소 RADIOISOTOPES 放射性同位元素

⋮ ⋮

RADIONUCLIDES 방사성동위원소 RADIOISOTOPES 放射性同位元素

\+ RADIOISOTOPES 방사성동위원소 RADIOISOTOPES 放射性同位元素

⋮ ⋮

放射性同位体 방사성동위원소 RADIOISOTOPES 放射性同位元素

\+ 放射性同位元素 방사성동위원소 RADIOISOTOPES 放射性同位元素

⋮ ⋮

<그림 17> 접근 어휘의 출력 예

4) 용어사전 출력

용어사전에서는 표제어와 동의어를 모두 출력하되 한국어, 영어, 일본어의 차례로 자모순 배열을 하였다. 각 용어 엔트리는 용어를 문단의 앞으로 약간 내어쓰고 진한 글씨로 처리하였으며, 수록 정보는 필드값이 있는 경우만 출력하였다. 용어의 정의는 문단을 나누어 출력하였고 맨 아래 줄에 개념ID와 용어ID를 출력하였다. 용어사전의 엔트리 출력 예를 그림 18에 제시하였다.

가스확산법 표제어. 명사. [이형] 가스 확산법 [분류] ICC:D5134

[정의](기체확산법) 동위원소 분리법의 하나. 분자량이 다른 가스(또는 증기)가 가는 구멍 또는 다공질의 격막을 통과하는 속도가 다른 점을 이용한 것. [출처] KND1996

C000000007 TK00000007

⋮

GASEOUS WASTES 표제어. 명사. 복수. [분류] ICC:D5131

[정의] 기체의 방사성 폐기물. 아르곤의 동위원소 41Ar과 같이 기체 자체가 불활성이며 화학적으로 처리가 곤란한 경우와 옥소의 동위원소 131I과 같이 기체로서도 반응, 흡수 등에 의해서 용액 중에 용해시킬 수가 있는 경우가 있어서 각각 처리법을 다르게 한다. [출처] KND1996

C000000143 TE00000143

⋮

アクチニウム元素 표제어. 명사. [분류] ICC:D5110

[정의] 원자번호가 90인 토륨 원소부터 103번인 로렌슘 원소까지 14개 원소의 총칭. 즉 actinoids 중 악티늄(Ac)을 제외한 원소의 총칭. 악티늄과 닮은 원소를 뜻하나 화학적으로는 란탄족을 닮았다. [출처] KND1996

C000000575 TJ00000575

<그림 18> 용어사전 엔트리의 출력 예

7. 결론

과학기술 분류표, 시소러스, 용어사전 등의 주요한 색인 및 검색 도구를 한국어, 영어, 일본어의 3개 언어로 통합 구축하고 활용할 수 있는 다기능, 다국어 과학기술 통합 개념체계의 개발 방안을 마련하고 원자력 분야를 대상으로 프로토타입 시스템을 구축하였다. 이 통합 개념체계의 응용 분야는 분류와 색인 등의 정보 조직, 온라인 검색과 필터링 등의 정보 검색, 자동번역 및 교차언어 정보검색 등의 다국어 처리 분야가 될 수 있다.

통합 개념체계가 앞으로 실제로 구축되어 이용이 활성화되려면 다음과 같은 후속 연구가 필요하다. 첫째, 패싯유형을 폭넓은 주제분야의 용어에 적용해서 검토해야 하며, 패싯유형 할당의 엄정한 원칙과 지침을 제시해

야 한다. 둘째, 분류표와 분류표 사이의 주제범주 매핑에 대해서 다양한 대응의 경우를 세밀히 검토하여 적절한 매핑 원칙을 마련해야 한다. 셋째, 프로토타입 시스템에서는 한국어, 영어, 일본어의 세 언어만 고려하였지만 통합 개념체계에 다른 언어를 추가로 수용할 경우에 대비해서 엔트리의 출력 형태를 개선할 여지가 있다. 넷째, 실제 개발은 통합분류표의 대분류를 기준으로 각 주제영역별로 진행하되 이전에 구축된 다른 주제분야의 요소를 활용하는 지침을 정리하여 제시해야 한다. 다섯째, 프로토타입 시스템을 실제 자료의 색인 및 검색에 적용하여 그 유용성을 검증하고 미비한 부분을 보완해야 한다.

참고 문헌

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

박영자, 송만석. 1992. "자연언어 처리를 위한 한국어 동사·명사의 개념 분류." 『제4회 한글 및 한국어정보처리 학술발표논문집』, 141-149.

설성수, 송충한. 1999. 『기초과학분야의 분야분류체계 개발연구』. 한국과학재단.

윤광진. 1989. 『의미소성의 계층구조에 의한 사전구성(I)』. 석사학위논문, 인하대학교 대학원 전자공학과.

정영미. 1997. 『지식구조론』. 한국도서관협회.

한국원자력산업회의. 1996. 『최신 원자력용어사전』. 한국원자력산업회의.

Aitchison, J. 1970. "A thesaurofacet : a multipurpose retrieval language tool." *Journal of Documentation*, 26(3): 187-203.

Calzolari, N. 1988. "The dictionary and the thesaurus can be combined." In M.W. Evens (ed.), *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, 75-96. Cambridge University Press.

CLS Framework: Listing of ISO 12620 Data Categories [WWW]. [cited 2001.7.15]. <http://www.ttt.org/clsframe/datcats.html>

Dag Hammarskjöld Library. 1995. *UNBIS thesaurus*. (English edition). United Nations.

General European Multilingual Environment Thesaurus(GEMET) [WWW]. [cited 2001.6.10]. <http://www.mu.niedersachsen.de/cds/documents/ws0996/gemet02.html>.

International Atomic Energy Agency. 1999. *INIS: Thesaurus*. Vienna: International Atomic Energy Agency.

International Organization for Standardization. 1999. *Computer Applications in Terminology - Macine-Readable Terminology Interchange Format(MARTIF) - Negotiated Interchange*. (ISO 12200:1999(E)). Geneva : International Organization for Standardization.

Japan Electronic Dictionary Research Institute. 1988. *Concept Dictionary*. EDR Technical Report TR-009.

Lenat, Doug, George Miller, and Toshio Yokoi. 1995. "CYC, WordNet, and EDR: critiques and responses." *Communications of the ACM*, 38(11): 45-48.

Melby, A. 1995. "E-Tif: An Electronic Terminology Interchange Format." *Computers and the Humanities*, 29: 159-165.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1990. "Introduction to

WordNet: an on-line lexical database." *International Journal of Lexicography*, 3(4): 235-244.

National Library of Medicine. 2001. *UMLS Information* [WWW]. [cited 2001.7.1]. <http://umlsinfo.nlm.nih.gov/>.

Office for Official Publications of the European Communities. 1995. *EUROVOC Thesaurus*, v.1: Permuted alphabetical version. - v.2: Subject-oriented version. - v.3: Multilingual version. Luxembourg.

Princeton University Cognitive Science Laboratory. *WordNet - a Lexical Database for English* [WWW]. [cited 2001.5.11]. <http://www.cogsci.princeton.edu/~wn/>.

Sager, J. C. 1990. *A Practical course in terminology processing*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Schmitz-Esser, Winfried. 1990. "Thesauri facing new challenges." *International Classification*, 17(3/4): 129-132.

Schmitz-Esser, Winfried. 1991. "New approaches in thesaurus application." *International Classification*, 18(3): 143-147.

Schmitz-Esser, Winfried. 1999. "Thesaurus and beyond: an advanced formula for linguistic engineering and information retrieval." *Knowledge Organization*, 26(1): 10-22.

Shiri, Ali Asghar, and Crawford Revie. 2000. "Thesauri on the Web: currents developments and trends." *Online Information Review*, 24(4): 273-279.

Soergel, Dagobert. 1996. "SemWeb: proposal for an open, multifunctional, multilingual system for integrated access to knowledge base about concepts and terminology." *Proceedings of the Fourth International ISKO Conference*, 165-173.

Soergel, Dagobert. 1999. "Enriched thesauri as networked knowledge bases for people and machines" Presentations from the Subject Analysis and Retrieval Working Group Conference Controlled Vocabulary and the Internet. [online]. [cited 2001.5.11]. <http://www.dtic.mil/cendi/presentations/cendisoergel.pdf>.

Text Encoding Initiative Consortium. 2001. *Guidelines for Electronic Text Encoding and Interchange*. (Preliminary XML edition).

Vossen, P., P. Diez-Orzas, and W. Peters. 1997. "The multilingual design of EuroWordNet." In *Proceedings of the ACL/EACL-97 workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Wake, Susannah, and Dennis Nicholson. 2001. "HILT - High-Level Thesaurus Project: building consensus for interoperable subject access across communities." *D-Lib Magazine* [online], 7(9). [cited 2001.9.16]. <http://www.dlib.org/dlib/september01/wake/09wake.html>.

Yokoi, Toshio. 1995. "The EDR electronic dictionary." *Communications of the ACM*, 38(11): 42-44.