94-b-4

# 백과사전 데이타베이스를 위한 색인시스템 설계

## The Design of Index System for Encyclopedia Database

추윤미, 최석두 이화여자대학교 문헌정보학과

Yoonmi Chu, Sukdoo Choi Dept. of Library and Information science, Ewha Women's Univ.

백과사전 데이타베이스의 효과적인 검색을 위한 색인시스템을 설계하였다. 여기에서는 항목에 대한 각종 속성정보와 본문정보를 모두 포함한 색인표제어파일을 작성하고, 각 항목에 대한 참조항목을 별도로 두지 않고 시소러스파일의 BT, NT, RT, UF를 사용하여 그 항목과 연관된 항목을 참조하도록 한다. 시소러스파일은 각 색인표제어에 부여한 주제분류기호(DDC, 또는 KDC)의 계층구조를 이용하여 자동생성한 후 색인자의 수작업을 거쳐 작성된다.

이 색인시스템을 통해 백과사전에 포함되어 있는 모든 정보를 이용한 다양한 접근이 가능하며 시소러스를 사용하여 관련항목을 브라우징을 할 수 있어 포괄적인 검색이 가능하다.

## 1. 서론

백과사전은 방대한 양의 데이타를 포함하고 있어서 데이타의 관리 및 유지가 어렵고 이용자의 검색에도 불편하다. 종래의 인쇄된 백과사전을 참조하기 위해서는 항목명의 가나다순으로 배열되어 있는 본책을 직접 찾거나 색인이 제공된다면 색인에서 찾고자 하는 항목의 소재위치를 찾는 것이 보통이다. 그러나 이러한 방법은 반드시 찾고자 하는 내용의 항목명을 알았을 때만 가능하며 어떤 주제에 대한 포괄적인 검색이 불가능하다.

이제까지 백과사전의 색인에 있어서의 문제점은 첫째, 내용이 방대함에 따라 색인의 항목수도 방대하다는 점, 둘째, 이용자의 연령층이 어린아이부터 노인에 이르기까지 매우 다양하고, 교육정도도 다양함에 따라 그 주제가 다양하다는 점, 셋째, 공동작업으로 일관성을 유지하기 어렵다는 점, 네째, 색인의 깊이와 대상이 다양하는 점 등이다(Clevleand, 1990).

본 논문에서는 이러한 백과사전 색인에 있어서의 문제점을 해결하는 방안으로 백과사전 데이타베이스를 기반으로 설계된 색인시스템을 제안하고자 한다. 백과사전 데이타베이스는 관계형 데이타베이스로 설계하였다.

## 2. 백과사전 데이타의 특성

백과사전의 각 항목들은 크게 항목정보와 본문으로 나눌 수 있다. 항목정보는 그 항목의 속성을 항목명, 속성분류(인명, 지명, 문화재, 작품, 단체, 사상, 문헌, 제도, 사건 등), 주제분류, 지리분류, 외래어, 한자, 약칭, 별칭, 참고문헌 등으로 분석해 낸 정보이다. 그러므로 항목정보는 정형적인 데이타형식을 지닌다. 반면에 백과사전의 본문은 전문(全文), 사진, 도판, 표 등의 다양한 데이타 형태로 구성되며 더 나아가 동화상, 음성까지 그 범위가 확대된다. 따라서 백과사전이 지닌 정형적/비정형적 데이타를 모두 포함하는 색

인이 필요하다.

백과사전은 존재하는 어떤 사물이나 개념에 대한 설명이다. 백과사전을 이용하는 이용자의 목적은 어떤 사물이나 개념에 대해 알고자 하는 것이며 이러한 욕구는 연관된 다른 정보를 브라우즈함으로써 더욱 충족시킬 수 있다. 그러므로 백과사전에서 브라우징 기능은 매우 중요하다. 이에 따라 색인표제어를 그 개념 관계에 따라 망구조로 조직하는 것이 필요하다. 또한 이렇게 조직된 색인을 이용한 하이퍼텍스트, 더 나아가 하이퍼미디어 검색기법이 절실히 요구된다.

## 3. 색인시스템

### 3.1 색인시스템의 구성

백과사전의 색인을 위한 시스템구성은 그림1과 같다.

1) 항목정보파일 : 항목에 대한 속성정보

2) 항목관련정보파일 : 본문을 구성하고 있는 각각의 자료(사진, 도판, 본문텍스트 등)의 속성정보

3) 원정보파일 : 사진, 도판, 본문텍스트, 표

4) 분류표 : 주제분류표는 DDC나 KDC를 기준으로 항목의 주제분류코드를 부여하기 위해 사용한다. 지리분류표나 시대분류표도 주제분류표에 준한다.

5) 참고문헌파일 : 각 원정보(사진, 도판, 본문, 표)의 참고문헌에 대한 정보

6) 저자파일 : 각 원정보(사진, 도판, 본문, 표)의 저자에 대한 정보

색인표제어파일 및 시소러스파일에 대해서는 후술한다.

### 3.2 색인파일의 구조

인쇄본 백과사전과는 달리 전자출판되는 백과사전은 공항목(항목에 해당하는 본문이 없이 다른 항목을 참조하도록 되어 있는 항목)을 별도로 설정, 관리할 필요가 없이 색인표제어파일에 색인표제어로 등록하고 참조할 항목만을 주어 링크시키면 된다. 또한 전자출판되는 백과사전에서는 그 항목의 소재위치를 따로 가질 필요가 없고 색인표제어파일과 항목정보파일을 링크시킬 수 있도록 항목ID를 가지고 있으면 된다. 또한 참조항목을 별도로 두지 않고 시소러스파일의 BT, NT, RT, UF를 사용하여 그 항목과 연관된 항목을 참조하도록 한다.색인표제어를 모두 등록한 후에 색인표제어간의 관계를 줌으로써 시소러스를 생성한다. 생성된 시소러스는 검색을 위한 용도로 사용된다. 다음의 그림2는 색인표제어파일과 시소러스파일의 구조와 예를 보여준다.

1) 색인표제어파일

① 색인ID : 색인표제어파일의 식별자.

② 색인표제어 : 색인어. 항목정보에서 항목명, 항목명의 별칭, 이칭, 약칭, 로마자, 영어, 한자, 한자한글음, 항목관련정보에서의 캡션제목, 본문텍스트에서 선정한 주요어 등을 색인표제어로 선정한다.

③ 한정어 : 동음이의어인 경우에 이를 구분하기 위하여 부기하는 말이다.

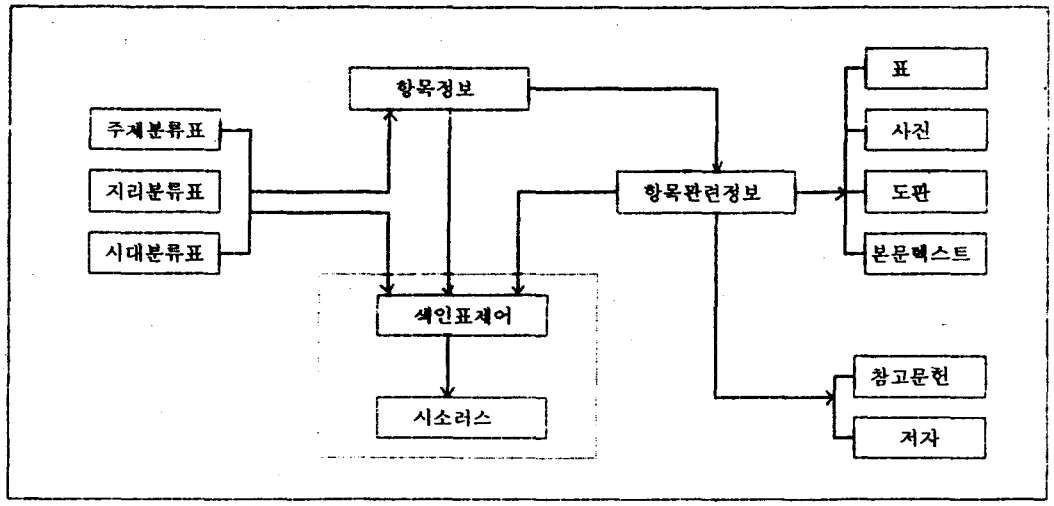


그림 1. 색인시스템의 구성도

예) 유노(여신)
유노(소행성)

④ 색인표제어구분 : 색인표제어가 어느 항목에서 추출되었는지를 나타낸다. 항목명 ,별칭 ,이칭, 약칭, 로마자, 영어, 한자, 한자한글음, 사진제목, 도판제목, 표제목, 본문중주요어 등이 있다. 사진제목, 도판제목, 표제목은 캡션제목을 말하는데 단어, 구에 상관없이 무조건 색인표제어로 사용하는 것을 원칙으로 한다.

⑤ 주제분류 : 각각의 색인표제어에 주제분류번호를 부여한다.

⑥ 지리분류 : 각각의 색인표제어에 대해 지리적 분류가 필요할 때 부여한다.

⑦ 시대분류 : 각각의 색인표제어에 대해 시대적 분류가 필요할 때 부여한다.

⑧ 속성분류 : 속성은 그 항목이 어떤 범주에 속하는 지를 나타낸다. 속성은 대체로 인물, 사건, 문화재, 지명, 사상, 문헌, 작품, 동식물, 제도, 단체, 기구 등으로 나눌 수 있다.

⑨ 항목ID : 그 색인어가 소속된 항목을 가리킨다.

2) 시소러스파일

① 색인ID : 색인표제어파일의 색인ID와 링크.

② 관계 : 색인ID와 지시ID간의 관계. NT, BT, RT, UF의 네 가지로 구분한다.

③ 지시ID : 하나의 색인표제어에 대하여 관계가 있는 색인표제어를 지시한다.

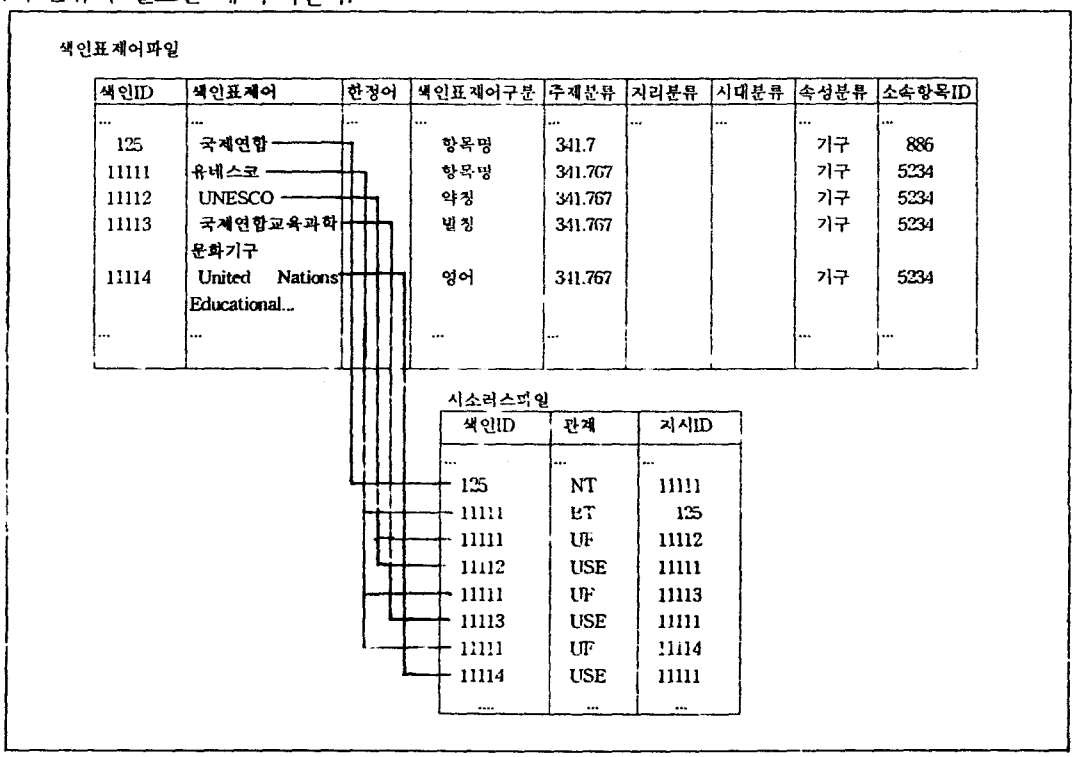


그림 2. 색인파일의 구조

### 3.3 색인표제어파일의 작성

1) 항목정보파일에서 표제어부에 해당되는 항목명, 영어명, 로마자명, 한자명, 학명 등을 색인표제어로 색인파일에 등록한다. 각 항목정보파일에서 각 항목에 부여된 주제분류, 지리분류, 속성분류를 색인표제어파일로 가져온다.

2) 항목관련정보파일에서 관련구분이 본문인 경우를 제외하고 제목을 색인표제어로 색인파일에 등록한다. 사진제목, 도판제목, 표제목 등은 주제분류, 지리분류, 속성분류를 지정한다.

3) 본문중에 있는 주요어를 추출한다. 추출방식에는 세가지가 있다.

① 수작업 : 색인자가 내용을 읽은 후 적절한 주요어를 지정하거나 주제어를 부여한다.

② 자동추출 : 컴퓨터가 내용을 분석하여 통계적기법이나 구문론적기법, 의미론적기법을이용하여 추출.

③ 반자동추출: 컴퓨터가 1차적으로 주요어를 추출한 후 색인자가 선정한다. 선정된 주요어는 본문텍스트에 마크업(Mark-up)을 하여 검색시에 주요어임을 알 수 있도록 한다(강현규 등, 1993). 항목정보파일을 검색하여 선정된 주요어와 같은 항목명이 있으면 본문텍스트에 마크업을 하여 그 항목과 링크시켜 준다.

### 3.4 시소러스 생성

시소러스는 정보자료를 색인시에 색인어를 통제해주는 역할과 필요한 정보자료를 검색할 때의 검색어의 선정을 통제해 주는 역할을 한다. 여기에서 시소러스는 모든 색인표제어를 하나의 망으로 조직함으로써 개념구조를 이용한 검색에 사용하고자 하는 목적으로 작성한다.

전자출판된 백과사전에서는 기존의 인쇄본에서의 직접적인 참조표시(see, see also)를 사용하지 않고 그 항목과 관련되는 상위어, 하위어, 유사어 등을 체계적으로 보여줌으로써 참조를 확대할 수 있다. 본 시스템에서 시소러스는 색인표제어와 주제분류기호를 이용하여 다음과 같이 생성한다.

1) 각 색인표제어에 주어진 주제분류기호의 계층관계에 따라 상위개념부터 BT, NT를 자동설정한다.

예) 화학 430
유기화학 437
→ 화학 NT 유기화학

2) 색인표제어구분이 별칭, 이칭, 약칭, 로마자, 영어, 한자한글음인 경우 그 항목명에 대하여 UF로 설정한다.

예) 장제스 한자한글음: 장개석
→ 장제스 UF 장개석

3) 같은 주제분류기호를 갖고 색인표제어구분이 항목명인 경우 RT로 설정하여 등록한 후 수작업으로 색인자가 조정한다.

4) 한 색인어에 대하여 시소러스파일의 레코드를 생성하면 시스템이 자동으로 지시항목에 대한 레코드를 생성한다.

예) 망원경 NT 적외선망원경
→ 적외선망원경 BT 망원경

망원경 RT 천문기계
→ 천문기계 RT 망원경

5) 색인자가 전체적으로 관계를 조정한다.

## 4. 검색 방법

위와 같이 구축된 색인시스템을 통해 다음과 같은 검색을 할 수 있다.

1) 항목을 전방일치, 완전일치로 검색

2) 본문중의 주요어에 대한 항목이 있을 때 그 항목으로 이동하는 하이퍼텍스트 검색

3) 불리언 연사자를 이용한 검색

4) 색인표제어구분 중 사진, 도판, 표제목 등을 이용한 제한검색

5) 항목중에서 지리에 관계된 것은 지리분류를 사용하여 검색

6) 항목의 지리분류에 의한 제한검색

6) 역사적 사건과 같이 시간과 관련된 것은 시대분류를 사용하여 검색

7) 주제분류표를 이용하여 주제로 검색

8) 항목이나 본문중의 주요어와 관련된 항목을 시소러스를 이용하여 확장 검색

9) 항목의 저자나 참고문헌을 이용하여 검색

## 5. 결론

위와 같은 백과사전 색인시스템을 구축함으로써 다음과 같은 효과를 기대할 수 있다.

첫째, 다양한 엑세스 포인트로 검색할 수 있다는 점이다.

둘째, 시소러스를 이용한 주제어의 검색과 개념 확장이 가능하여 포괄적인 검색을 할 수 있다.

세째, 검색시 브라우징이 편리하다.

네째, 각 항목의 관련정보는 재사용이 가능하여 이차자료의 기능을 할 수 있다.

전자출판이 발전되면서 백과사전 자료의 형태도 영상, 음성자료와 같은 멀티미디어데이타를 포함하게 되었다. 그러나 아직까지 현재의 기술로는 멀티미디어 데이타의 속성을 분석하여 색인할 수 있는 색인기술에 미치지 못하고 있다. 이미지 처리나 시그널 처리의 기술이 인식의 범위까지 가지 못했기 때문이다(박세영, 1994). 따라서 캡션의 내용을 분석하여 이에 알맞는 색인어를 부여해야 한다. 앞으로 단순히 문장에 출현한 단어를 대상으로 색인하는 것이 아니라 이러한 설명문이나 본문의 색인을 의미분석을 통한 색인어의 부여가 있어야 한다. 이를 위해 백과사전 데이타의 특성 파악과 언어적 기법이 도입되어야 할 것이다.

## 참고문헌

Cleveland, B. D, Ana D. Cleveland(1990). Introduction to Indexing and Abstracting. Libraries Unlimited.

강현규, 이창열, 박세영(1993). "백과사전 검색 시스템의 설계 및 구현." 한국정보과학회 가을 학술발표논문집 Vol. 20, No. 2, pp 1167-1170.

박세영(1994). "멀티미디어 정보검색에서의 한국어 정보처리". 한국정보과학회지 Vol.12, No 8, pp 60-65.