정보시스템의 유니코드 기반 한자검색 지원

Support on Ideograph characters search of unicode based Information System

윤 소 영 (SoYoung Yoon)*

초 록

현재 유니코드 CJK 한자코드는 부수 기준 배열방식을 따르고 있어 한자의 한글음가를 기준으로 하는 우리의 문자생활 방식과 차이가 있으며, 우리나라 고유한자나 동형이음어, 이두문자, 그리고 이체자 관계 등을 모두 수록하고 있지 않아 정보시스템에 그대로 적용하기에는 무리가 있다. 따라서 유니코드 기반 정보시스템의 정확한 한자표현 및 한자검색을 위해서는 한자를 포함하는 자료에 대한 정확한 이해를 바탕으로 여러 가지 지원방안을 마련해야 한다. 이러한 측면에서 역사분야 정보검색시스템에서는 한글음가 및 한국에서만 사용되는 동형이음어 처리를 위한 한자-한글음가 사전, 본래 한자의 음가와 다르게 읽히는 한자를 위한 특수용어사전, 이형자와 이체자를 위한 이체자사전, 그리고 유니코드 CJK 통합한자에 등록되어 있지 않은 한자를 위한 신출한자목록을 지원하고 있다.

ABSTRACT

Unicode Han ideograph character set differed from the our principle of the phonetic value ordering in that it followed the principle of KangXi radical-stroke ordering of the characters. Therefore, information system should support ideograph search on precise analysis of materials which consist of korean character (hangul) and ideograph character (hanja). History Information system has been maintaining Hanja(Chinese Character) to Hangul Dictionary, Terminology Dictionary for composition, borrowing, non-ideographic principles, Variant Forms Dictionary, and Recently discovered Chinese Characters List.

키워드 : 유니코드, 한자검색, 문자표현, 한글음가, 이체자
Unicode, ideograph character search, character presentation, hangul phonetic value, variant forms of chinese character

* 국사편찬위원회 사료연구위원 (syoon@moe.go.kr)

1. 서론

표준 전쟁을 거쳐, 컨텐츠 전쟁이라고 일컬어지고 있을 만큼, 지금 우리의 21세기는 정보의 효율적인 이용을 최우선으로 하고 있다. 이용자를 위한 정보시스템 구축의 화두가 시스템이 대상으로 하고 있는 컨텐츠 자체의 효과적인 표현과 교환으로 옮겨지고 있다고 할 수 있다. 따라서 높은 수준의 IT와 결합하여 컨텐츠 산업에서 그 가치를 확산 시킬 수 있는 기본 전제조건으로서 문자표현에 관한 연구, 컨텐츠 기획, 구축 및 이용에 대한 지속적인 노력이 필요한 상황이다. 문화적 측면에서 다루어지고 있는 우리나라 문자에 대한 표준화가 경제적, 산업적 측면에서 그 가치를 확대하고 있다. 세상의 모든 문자가 코드화 되고 이 문자코드를 누구나 공통적으로 사용할 수 있게 표준화된다면 더 이상 정보시스템의 문자표현이나 데이터 처리에 있어 고민할 필요가 없게 될 것이다. 미래 지식정보 산업의 주도를 위해서는 자국의 문자에 대한 표준화가 반드시 선행되어야 한다. 유니코드에 자국의 문자가 등록이 되면 유니코드 기반 소프트웨어를 통해 전 세계 어디에서나 자국의 컨텐츠를 자유롭게 사용 할 수 있게 될 것이다.

1990년대 초 컴퓨터계의 다국적 기업들이 중심이 되어 결성한 유니코드(UniCode) 컨소시엄은 데이터 처리의 장애 요인이 되어 온 문자코드 문제를 근본적으로 해결하고자 하는 취지에서 문자코드의 세계적 표준화를 제안하였다. 국제표준코드인 유니코드는 1996년 v.2.0이 발표되어 일부 운영체제의 표준코드로 채택되었으며, 2006년에는 v.5.0을 거쳐 2007년 현재는 5.1.0 베타버전이 발표되었다. 향후 대부분의 운영체계나 응용소프트웨어가 유니코드를 표준코드로 채택, 확대될 것이다. 그러나 이러한 문자코드의 바벨탑은 현실적으로는 각 국의 문자생활에 맞도록 문자표현을 적용해야 하는 상황에서 많은 문제점을 나타내고 있다.

정보를 표현하기 위한 문자코드의 표준화와 더불어 대상 자료의 특성에 대한 정확한 분석을 통하여 컨텐츠의 특성을 반영하는 검색 지원 개발 및 활용 측면에서 계속적인 노력을 필요로 하고 있다. 현재 유니코드 표준 CJK 통합한자는 한자문화권에서 공통적으로 이용되는 방식을 따라 부수, 획수를 기준으로 문자를 배열하고 있다. 이에 더해 새로운 추가되는 문자는 기존 문자코드 다음에 추가되고 있기 때문에, 이 또한 정확하게 획수를 기준으로 배열되어있다고 할 수 없다. 우리나라는 한글과 마찬가지로 한자도 음가를 기준으로 배열하는 문자표현 체계를 이용하고 있기 때문에 유니코드 기반 정보시스템에서도 검색을 수행할 때 이를 고려해야 한다. 또한 우리나라에서만 사용되는 동형이음어 처리 등에 대한 지원도 있어야 한다.

인문과학 분야, 특히 역사, 문화 분야와 같이 한자 및 옛한글 등의 문자가 많이 포함되어 있는 컨텐츠를 대상으로 하는 원문검색시스템을 구축할 경우에는 이러한 측면에서 더

많은 주의가 필요하다. 동일 한자이나 다양하게 읽히는 경우, 그 자형이 다른 이형자나 이체자, 그리고 한자를 차용하여 우리말을 표현한 이두문자나 원래 한자의 음가와 전혀 다르게 읽히는 특수용어 등을 지원해야 한다. 이렇게 다양한 문자를 포함하는 컨텐츠의 구축과 이용의 전제 조건은 다음과 같다고 할 수 있다.

첫째, 원문은 있는 그대로 입력할 수 있어야한다. 원자료의 문자는 변형이나 대체 없이 그대로 입력하여야 그 자료의 의미를 제대로 전달할 수 있다.

둘째, 원문은 있는 그대로 볼 수 있어야한다. 원자료의 문자는 원상태 그대로 보여주어야 그 의미를 제대로 파악할 수 있다.

셋째, 시스템의 문자표현은 제한이 없어야한다. 시스템에 따라 표현할 수 있거나 없는 문자가 있어서는 안된다.

넷째, 검색과 문자표현은 상호 독립적이어야한다. 시스템에서 지원하는 검색엔진이나 어플리케이션이 원자료의 문자표현에 영향을 주어서는 안된다.

표준 유니코드 기반으로 문자코드를 변환한다 하더라도 유니코드의 한글, 한자코드의 여러 가지 제약 특성상 실제 정보시스템에서 문제점들이 나타난다. 국제 표준인 유니코드 기반으로 우리나라 자료의 특성을 반영한 한자검색을 지원하도록 정보시스템을 개발, 적용하고 공유한다면 컨텐츠의 경쟁력을 한층 더 높일 수 있을 것이다. 이러한 관점에서 본 연구에서는 유니코드 적용에 따르는 한자, 한글 관련 문제점을 해결하기 위하여 역사분야에서 개발, 활용 중인 한자한글음가 사전, 특수역사용어사전, 이체자사전, 그리고 신출한자목록 등을 그 해결방안으로 제안하고자 한다.

2. 유니코드기반 문자표현

2.1 유니코드의 한글코드

소프트웨어 업체들을 중심으로 다양한 언어를 위한 문자표현 코드체계를 만들려는 의도를 가지고 유니코드 컨소시엄을 결성하였다. 이와 동시에 1980년대 초반부터 국제표준기구를 주축으로 하여 문자표현을 위한 코드체계가 새로 작성되기 시작하였다. 1991년부터 유니코드 컨소시엄의 참여로 급속히 발전을 거듭한 ISO/IEC 10646 (UCS; Universal Multiple-Octet Coded Character Set) 혹은 유니코드라고도 불리는 새로운 코드체계가 전세계적으로 널리 보급되어 사용되고 있다.

국내에서도 1985년부터 컴퓨터 제조업체마다 각기 다르게 사용하고 있던 한글코드를 통

일시키기 위해 한글 표준코드를 연구하기 시작하여 1987년에 행정전산망용으로 ISO-2022 코드체계를 확장하여 만든 'KS C 5601(1998년 KS X 1001로 개칭)' 완성형 한글코드를 발표하였다. 이 코드는 현대 한글 2,350자, 한자 4,888자, 특수문자 1,128자, 그리고 470자의 사용자 영역으로 구성되었으나 옛한글은 제외되었다. KS C 5601의 제정으로 한글코드의 표준화를 이루었지만 반드시 필요한 상당수의 글자가 빠져 있고, 옛한글을 입력할 수 없는 등 우리 문자생활에 불편을 초래하였다. 이러한 문제점을 보완하기 위해 1991년에 한글 1,930자와 옛한글 1,677자를 완성형으로 추가하여 'KS C 5657(2001년 KS X 1002로 개칭)'을 제정하게 되었다. 완성형으로 추가된 옛한글은 음절로만 구성된 것으로, 고유어를 표기하기 위해 필요한 최소한의 글자만 수록하였다. 표현 가능한 글자가 제한적이었으나, 처음으로 옛한글 표준화를 시도하였다는 데에 의의가 있다.

1991년까지는 한글이 KS C 5601 표준을 중심으로 유니코드에 등록되어 있었기 때문에 제한된 숫자의 한글만을 사용할 수 있었다. 이후 1992년 ISO/IEC JTC1/SC2 국제회의를 유치한 한국 대표단은 BMP(Basic Multilingual Plane; 기본 다국어 판)에 현대한글과 당시까지 알려진 옛한글을 모두 포함하여 조합형 한글 자모 240자(초성 90자, 중성 66자, 종성 82자,채움 2자)를 추가하여 본격적으로 한글코드를 적용할 수 있게 되었다. 이후 1995년에 많은 전문가들의 노력에 힘입어 중국의 반대에도 불구하고 유니코드 2.0판에 현대한글 11,172자를 완성형 형태로 BMP에 추가하여, 효율적인 한글코드 사용이 가능하게 되었다(안대혁, 박영배. 2007).

.옛한글 자모 240자 : U+1100～U+11F9(Hangul Jamo) 영역에 초성 90자, 중성 66자, 종성 82자, 채움 2자로 구성

.한글 호환용 자모 94자 : U+3131～U+318E(Hangul Compatibility Jamo) 영역에 초.중.종성 구별 없이 수록

.현대 한글 11,172자 : U+AC00～U+D7A3(Hangul Syllables) 영역에 초성 19자, 중성 21자, 종성 27자로 조합할 수 있는 현대 한글 11,172자를 완성형으로 수록

'조합형 한글 자모'에 의한 조합형 한글의 표현은 한글 초성·중성·종성을 이용하여 한글을 표현하는 것으로 한글을 자유롭게 표현할 수 있다는 면에서 학계에서 선호하고 있는 방법이다. 다만 한 음절을 표현하는데 사용되는 메모리가 두 세배이상 요구되고, 한 음절의 길이가 일정하지 않기 때문에 일반적인 소프트웨어 처리에서의 효율이 떨어져 특별한 용도에 국한되어 사용되고 있다. 그러나 실생활에서 컴퓨터로 한글을 사용하는 할 때 현대 한글 음절 11,172자와 한글자모로도 충분하다고 할 수 있겠으나, 옛한글을 포함하는 자료와 언어처리를 위한 자모 조합에 의한 한글표현은 꼭 필요하게 된다.

옛한글은 240자의 옛한글 자모를 조합하는 방식을 사용하여 처리하도록 하였는데, 즉

현대 한글은 완성형으로, 옛한글은 조합형으로 구현할 수 있도록 하였다. 옛한글로 표현 가능한 글자의 범위는 예전에 비해 훨씬 넓어지게 되어 고유어뿐만 아니라 동국정운식(東國正韻式) 한자음은 물론, 한어(漢語), 몽고어(蒙古語), 일본어(日本語)를 한글로 표기한 글자, 개화기 외래어 표기를 위한 글자 등도 대부분 표현할 수 있게 되었다. 그러나 이것만으로는 옛한글 자료를 컴퓨터에서 완전하게 처리한다는 것은 불가능하다. 유니코드는 현대 한글을 완성형으로 옛한글은 조합하여 사용하도록 하고 있으나, 현대 한글 자모와 옛 한글 자모의 배열이 분리되어 있다. 유니코드는 한글 자모 영역 안에서 자모를 초・중・종성으로 나누어 배열하였으며 초・중・종성 안에서는 현대 한글 자모를 먼저 가나다순으로 먼저 배열한 다음 다시 옛 한글에만 쓰이는 자모를 가나다순으로 배열하였다. 유니코드로 한글을 표현할 경우 첫가끝 코드의 경우 부호자리가 1100～11FF 영역에 배열되어 있고, 완성형의 경우 AC00～D7A3 영역에 배열되어 있으므로 첫가끝 코드로 나타낸 한글이 완성형 한글표현보다 항상 앞에 오게 된다. 따라서 이러한 배열 순서가 현대 한글과 옛 한글의 순서를 정렬함에 있어 문제가 되며 이는 고문헌 자료에 대한 데이터베이스화 작업이 어려워질 수 있다.

유니코드 정규화 원칙에 따라 옛한글을 조합할 경우에도, 일부 글자가 완성자를 포함한 글자로 변환되는 문제가 발생하는데, 이는 한글을 입력하는 방식으로 완성형과 조합형을 동시에 사용하면서 발생하는 문제이다. 한글 표현을 위한 표준적인 문자코드를 제정할 때, 그리고 그 연장선상에서 유니코드에 한글 문자코드를 포함시킬 때 한글의 생성원리에 대한 원천적인 고민보다는 입력, 처리의 편리성을 더 우선시하여 빚어진 결과라 할 수 있다(안대혁, 박영배. 2007). 그러나 <옛한글 자모 확장목록> 117자가 BMP 영역에 최종 등록됨으로써 이러한 문제가 어느 정도 해결 되어, 옛한글로 이루어진 자료의 입출력 및 검색의 토대가 마련되었다.

지난 4월 독일 프랑크푸르트에서 열린 국제 표준화 기구인 ISO/IEC 의 JTC1/SC2/WG2 제50차 회의에서 한국 측이 제안한 <옛한글 자모 확장목록> 117자가 유니코드 BMP 영역에 추가로 등록되었다. 이번에 등록된 옛한글 자모는 대부분 겹자모로 광범위한 전거자료 조사를 바탕으로 작성된 것이며, 기존에 등록된 옛한글 자모와 구별하여 <옛한글 자모 확장목록>이라 한다. <옛한글 자모 확장목록> 117자의 BMP 상의 영역은 다음과 같다.

<초성: 34자>

Hangul Jamo: U+115A～U+115E

Hangul Jamo Extended-A: U+A960～U+A97C

<중성: 28자>

Hangul Jamo: U+11A3～U+11A7

Hangul Jamo Extended-B: U+D7B0～U+D7C6
<종성: 55자>
Hangul Jamo: U+11FA～U+11FF
Hangul Jamo Extended-B: U+D7CB～U+D7FB

2.2 유니코드의 한자표현

한중일 통합한자(CJK Unified Ideographs)는 유니코드에서 한국, 중국, 일본에서 사용하고 있는 한자를 하나로 통합하여 유니코드에 등록하면서부터 세 나라의 영문 이니셜을 따서 'CJK 통합한자'라고 불렀다. 일부에서는 베트남에서 사용되던 한자가 추가된 이후부터는 'CJKV 통합한자'라고도 부른다.

유니코드 설계 당시 65,536개의 문자영역 속에 세계의 문자를 모두 수록하고자 하였는데, 한자를 이 한정된 문자영역안에 통합하면서 여러 가지 문제가 발생하게 되었다. 같은 의미를 같은 목적으로 표현하는 문자는 가능한 같은 코드값에 할당한다는 유니코드의 방침에 기초하여 각국의 미묘한 자형의 차이는 무시하고 비슷한 한자를 통합하였다. 그러나 기존의 코드 체계에서 구별되고 있는 한자는 별도의 코드값을 부여하고, 번체자와 간체자 등 명확하게 자형이 다른 경우에도 별도의 코드값을 부여하였다.

그 뒤 유니코드의 코드 할당가능 영역이 기본언어판(BMP) 외에 16개의 언어판이 추가 확대되어 한중일 통합한자도 기본언어판의 20,902자 외에 확장 A(Ext.A) 6,582자, 확장 B(Ext.B) 42,711자, 확장 C(Ext.C) 4,219자 등 계속 추가되고 있다. 이렇게 다량의 한자가 추가되는 과정에서 통합원칙을 지키지 않아 완전히 같은 글자가 복수의 코드값으로 등록되거나 최초의 20,902자에서 통합되었다가 확장 B에서 다시 등록되는 등 문제점을 안고 있는 것도 사실이다.

	V1.0.0	V1.0.1	V1.1	V2.0	V2.1	V3.0	V3.1	V3.2	V4.0	V4.1	V5.0
Alphabetics, Symbols	4,734	4,728	6,290	6,491	6,493	10,210	11,798	12,753	13,973	15,117	16,486
Han (URO)		20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902	20,902
Han (URO Extension)										22	22
Han Extension A						6,582	6,582	6,582	6,582	6,582	6,582
Han Extension B							42,711	42,711	42,711	42,711	42,711
Han Compatibility		302	302	302	302	302	844	903	903	1,009	1,009
Subtotal Han		21,204	21,204	21,204	21,204	27,786	71,039	71,098	71,098	71,226	71,226
Hangul Syllables	2,350	2,350	6,656	11,172	11,172	11,172	11,172	11,172	11,172	11,172	11,172
Graphic Characters	7,084	28,282	34,150	38,867	38,869	49,168	94,009	95,023	96,243	97,515	98,884
Format Characters	12	12	18	18	18	26	131	133	139	140	140
Graphic + Format	7,096	28,294	34,168	38,885	38,887	49,194	94,140	95,156	96,382	97,655	99,024
Controls	65	65	65	65	65	65	65	65	65	65	65
Private Use	5,632	6,144	6,400	137,468	137,468	137,468	137,468	137,468	137,468	137,468	137,468
Total Assigned	12,793	34,503	40,633	176,418	176,420	186,727	231,673	232,689	233,915	235,188	236,557
Surrogate Code Points				2,048	2,048	2,048	2,048	2,048	2,048	2,048	2,048
Noncharacters	2	2	2	34	34	34	66	66	66	66	66
Total Designated	12,795	34,505	40,635	178,500	178,502	188,809	233,787	234,803	236,029	237,302	238,671
Reserved Code Points	52,741	31,031	24,901	935,612	935,610	925,303	880,325	879,309	878,083	876,810	875,441

그림 1. 유니코드 버전별 문자코드 배정(The Unicode Consortium. 2007. p.1101)

우리나라가 한자 문화권에 속해 있기 때문에 많은 자료에 한자가 통용되면서 나타나는 문제도 적지 않다. 특히, 인명용 한자의 경우에는 표준 문자코드에 없는 글자가 많아 문제가 더욱 크게 부각되고 있다. 호적 전산화 시스템에서 한자를 등록하는 과정에서 표준 문자코드에 채택되지 않은 한자를 사용하는 사람의 경우에는 이름을 컴퓨터에서 기록하지 못하게 된다. 이러한 문제를 해결하기 위해서 국가차원에서 인명용 한자를 체계적으로 정리하여 해당 한자를 국제 표준에 등록하기 위한 노력을 진행하고 있다(박종우. 2007). 지금까지 우리나라가 제안하여 유니코드 표준에 수용된 한자는 CJK 통합한자 약 7만자 중에서 총 18,065자로 다음과 같다.

- BMP 한자
  - -CJK 통합한자 영역 : 15,390자
  - -CJK 통합한자 확장영역 Ext.A : 1,836자
  - -CJK 호환영역 : 268자
- CJK 통합한자 확장영역 Ext.B : 166자
- CJK 통합한자 확장영역 Ext.C : 405자

KSC5601-1987 표준코드에서 한자는 한자음 순으로 배열하고, 동음자내에서는 부수별 획수 순으로 배열하기 때문에 우리가 일반적으로 사용하는 한자음 순으로 이루어졌다. 그러

나 유니코드에서의 한자는 부수 순으로 배열되어 있기 때문에 정렬에 문제가 생긴다. 예를 들어 刻(각)과 街(가)의 경우 2획의 刂(도)가 부수인 刻(각)이 6획의 行(행)이 부수인 街(가)보다 선행되어 정렬된다.

그리고 두 가지 이상의 음을 가지는 한자의 배열의 경우에도 문제가 될 수 있다. 예를 들어 한자 '수레 차(車)'의 경우 '수레 거'라고 읽을 수도 있다. 이러한 경우 기존의 KSC5601 완성형에서는 한자의 정렬과 해당 한자를 발음에 따라 한글로 변환하는 데 있어서 모호성을 제거하기 위해 동형이음 한자를 모두 코드체계에 수용하였다. 반면에 유니코드에서는 대표적인 음인 수레 차(車)만을 수용하고 수레 거는 기존 KSC5601 표준과의 변환을 위해 배정된 유니코드 호환한자 영역에 배치하였다.

한자와 관련한 문제는 유니코드에 한자코드가 부수, 획수 기준 배열방식을 따르고 있으며, 유니코드 한자에 등록되어 있지 않은 우리나라에서만 사용되는 동형이음어가 존재하는 등 우리의 문자생활 방식과 차이가 있어 아직은 유니코드 자체만으로 해결하기에는 무리가 있어 보인다. 따라서 정보시스템에서 한자표현 및 검색 문제는 사전 테이블을 지원하는 방식으로 해결하는 것이 현실적인 대응방안이라 할 수 있다.

3. 한자검색 지원

한자가 가지고 있는 특성, 즉 음가가 2개 이상이라든지 특정 어휘로 쓰일 때 음가가 달라진다든지 하는 특수성으로 인해, 일반 이용자들은 물론이고 한자에 대한 전문지식을 갖춘 전문가들도 자료를 검색하고 활용하는데 어려움을 겪고 있다.

한자로 구축되어 있는 데이터베이스를 보다 편리하게 이용할 수 있도록 한글음 검색기능을 구현하였으나 하나의 한자가 2개 이상 음가를 갖는 경우가 많아 잉여 자료가 지나치게 많이 검색됨으로써 원하는 자료를 찾기 위해 다시 한 번 노력을 기울여야 하는 문제가 발생하였다. 한자의 특수성으로 인한 검색의 혼란을 최소화하기 위한 지원방안으로 현재 '한국역사정보통합시스템'을 비롯한 역사분야 정보검색시스템에서는 한자-한글 음가사전, 특수용어사전, 이체자사전 등을 공동으로 구축하여 제공하고 있다. 이와 더불어 신출한자 정리를 통해 한자 자료의 표현 및 이용에 불편함을 최소화하고 있다.

3.1 한자-한글음가 사전

한자는 한국, 중국, 일본, 대만 등에서 주로 사용하기 때문에 유니코드에서는 각 국가별로 사용되는 한자를 단일화 원리에 의해서 부수와 획수에 맞춰 배열하고 있다. 한자를 부수와 획수 순으로 배열하였기 때문에 우리의 문자생활방식에 맞는 한국어 음절순으로 정렬

하기 위해서 한자에 대한 한글 음절 순 테이블을 작성하여 적용하였다.

한자에 대한 한글음인 음가는 대부분 단음절로 되어 있으나 2자 이상으로 읽히는 예도 있다. 한자 데이터에 대해 한글음을 색인을 할 때 적용하는 음가로서 음가 사전테이블 항목 중 '색인 음가'에 있는 음은 모두 대표음이 된다. 한자-한글음가 사전의 주목적이 한자 데이터베이스 이용을 위한 가장 최적의 음가를 선정하는 데 있기 때문에, 국내에서 간행된 사전과 문화관광부 학술용역 과제로 선행된 표준음가 연구조사 결과를 적극 수용하여 대표음가를 선정하도록 하였다(강대걸. 2005).

표 1에서 예시한 유니코드 한자영역의 한중일 통합한자에 대해 대표음가를 확정하고 다중음가로 색인할 한자를 선정하고, 실제 데이터를 대상으로 검색을 수행한 결과를 반영하고 특수역사용어 사전테이블과 연동하여 수정사항을 반영하여 최적화 과정을 거쳐 한자-한글음가 사전테이블이 완성한다. 한자-한글 음가 사전 테이블은 그림 2에서 보는 바와 같이 자형, 유니코드, 그리고 대표음을 나열하는 구조로 구성된다.

표 1. 유니코드 한자 영역

구분	코드 영역	문자 수	누적 문자수	한글 음가
한중일 통합한자 영역 BMP(Basic Multi-lingual Plane)	U+4E00(一) ~ U+9FA5(鿥	20,902	20,902	• 유니코드 표준음으로 8,760개 음가 제안
한중일 통합한자 확장 A 영역	U+3400(㐀 ~ U+4DB5(䶵	6,582	27,484	• 유니코드표준음으로 제안한 음가 없음
한중일 통합한자 호환 영역(KSX)	U+F900(豈) ~ U+FA0B(廓)	268	27,752	• 우리나라 문자표준인KSC 5601에서 배정한 코드 • 표준음으로 288개 음가 제안
한중일 호환 영역(Big 5)	U+FA0C(兀) ~ U+FA0D(倕	2	27,754	• 제안음가 없음 • 중국 제안 한자
한중일 호환 영역(IBM 32)	U+FA0E(塚 ~ U+FA2D(鶴)	32	27,786	• 제안음가 없음 • U+FA2E, U+FA2F코드에는 배정된 문자 없음
한중일 호환 영역(JIS X)	U+FA30(侮) ~ U+FA6A(頻)	59	27,845	• 제안 음가 없음 • 일본 제안 한자

	A	B	C	D	E	F	G	H
1	자형	코드	대표음1	대표음2	대표음3	대표음4	대표음5	대표음6
737	姳	U+36E4	리	이				
738	媶	U+36E5	탑					
739	婻	U+36E6	람	남				
740	媚	U+36E7	맘					
741	媄	U+36E8	강					
742	嫥	U+36E9	축					
743	嫨	U+36EA	엄					
744	嫝	U+36EB	석					
745	嫮	U+36EC	륙	육				
746	媎	U+36ED	석					
747	嫈	U+36EE	수					
748	嫚	U+36EF	면	반	찬			
749	媚	U+36F0	혼					
750	嫟	U+36F1	외					
751	嫧	U+36F2	발					
752	嫈	U+36F3	예					
753	嫟	U+36F4	뇌					

그림 2. 한자-한글음가 사전

3.2 특수용어사전

특수용어는 한자의 특수성에서 기인하는 것으로서 음가 색인에서 제외된 음가를 포함하고 있는 용어를 뜻한다. 보다 넓게는 '한국역사정보시스템'에서 웹을 통해 제공하는 자료에 출현하는 용어로서, 동일한 대상이나 사건을 가리키는 용어임에도 다르게 표기한 것을 말한다. 예를 들면, 3.1운동(三一運動), 정1품(正一品)처럼 한자를 숫자로 표기하는 용어 등을 포함한다.

특수용어 사전은 다음과 같은 과정을 거쳐 구축하였는데, 첫째, 다양한 관련 사전에서 특정 어휘로 사용될 때, 대표음가와는 다른 음가를 갖는 용어를 추출하고 이를 질의어 확장 테이블 형태로 정리한다. 둘째, 근현대사 자료에서는 숫자가 포함된 용어들을 한자, 한글 외에 아라비아 숫자를 사용하여 표현되어 있기도 하므로 이러한 용어들도 수집하여 정리한다. 마지막으로, 실제 데이터를 대상으로 검색을 수행한 결과를 반영하고 한자-한글 음가 사전테이블과 연동하여 수정사항을 반영한다(강대걸. 2005).

2005년말 현재 5,680개 용어에 대해, 2,631개의 질의어-확장어 세트를 수록하고 있으며,

2006년에는 불교관련 자료가 한국역사정보통합시스템에 연계되면서 불교에서 쓰이는 특수용어들도 이 사전에 통합되어 특수 불교용어도 이용가능하게 되었다. 특수용어사전 테이블은 그림 3에서 보는 바와 같이 한자음, 연계한글음, 대표어(대표용례), 빈도수 등으로 구조로 구성한다. 특수용어사전 테이블을 검색엔진에 적용하여 색인을 작성하고, 그에 대한 검색을 수행한 결과는 그림 4에서 볼 수 있다. 이 예에서는 평민의 아내를 뜻하는 '조이(召史)'에 대한 검색결과를 나타내고 있다. 이두를 사용하는 한자자료에 대한 검증을 위해 구성한 이두자료읽기사전 테이블의 모습은 그림 5에서 볼 수 있다.

	A	B	C	D	G	H	I	J	L
1	한자음연계	연계음가	대표어	대표어빈도수	검색어	역통검색결과	Naver검색결과	국편검색결과	확장1
2192	蛻	열	蜘蛻	2	유열				蜘蛻
2193	召	조	干三召二	2	干三召二	0	-	0	간삼조이
2194	召	조	干三召二	2	간삼조이	0	-	1	干三召二
2195	召	조	大召	2	대조	0	-	0	大召
2196	召	초	召兒	3	초의	971	-	4857	召兒
2197	召	초	召兒	3	召兒	1	-	3	招兒
2198	召	초	召兒	3	招兒	1	-	2	招兒
2199	繅	조	玉繅	2	玉繅				옥조
2200	繅	조	玉繅	2	옥조				玉繅
2201	召史	조이	召史	2	조이	0	-	0	召史
2202	召史	조이	召史	2	召史	0	-	0	조이
2203	率	수	副衛率	2	부위수				副衛率
2204	率	수	副衛率	2	副衛率				부위수
2205	率	수	衛率	2	위수				衛率
2206	率	수	衛率	2	衛率				위수
2207	率	수	倡率	2	창수				倡率
2208	率	수	倡率	2	倡率				창수
2209	衰	최	墨衰	4	묵최				墨衰

그림 3. 특수용어사전

그림 4. 특수용어사전을 적용한 검색결과

	A	B	C	D	E	F	G
1	한자음연계	대표어		대표어 빈도수	검증	확장 어휘 수	검색어
170	이두	白乎所	3	7	TRUE	7	사온 바
171	이두	白乎所	3	7	TRUE	7	여짜온 바
172	이두	白乎所	3	7	TRUE	7	옵신 바
173	이두	白乎所	3	7	TRUE	7	ᄉᆞᄅᆞᆫ바
174	이두	白乎所	3	7	TRUE	7	ᄉᆞᄅᆞᆫ바
175	이두	白乎旀	3	4	FALSE	7	白乎旀
176	이두	白乎旀	3	4	FALSE	7	사오며
177	이두	白乎旀	3	4	FALSE	7	ᄉᆞᄒᆞ며
178	이두	白乎旀	3	4	FALSE	7	ᄉᆞᄒᆞ며
179	이두	不喩在	3	5	FALSE	7	不喩在
180	이두	不喩在	3	5	FALSE	7	아닌
181	이두	不喩在	3	5	FALSE	7	아닌 것
182	이두	不喩在	3	5	FALSE	7	아닌디견
183	이두	不喩在	3	5	FALSE	7	아닌지견
184	이두	順可只	3	4	FALSE	7	로부터
185	이두	順可只	3	4	FALSE	7	順可只
186	이두	順可只	3	4	FALSE	7	순함직
187	이두	順可只	3	4	FALSE	7	조차도
188	이두	順音可	3	3	FALSE	7	順音可
189	이두	順音可	3	3	FALSE	7	순홈직

그림 5. 특수용어사전-이두자료읽기사전

3.3 이체자 사전

유니코드내에서 이체자 관계에 있는 4,380종 9,309자로 확대하고, 여기에 민족문화추진회(현, 한국고전번역원)의 한국문집총간 정보화사업 수행 과정에서 수집된 이체자 자료를 추가하여 이체자 사전 테이블을 구성하였다. 유니코드3.0(Extension A 영역 한자 포함) 27,484에 대한 이체자 정보에 대해 각각 대표자와 이체자(별자), 그리고 각 대표자와 이체자에 대한 이형자로 구분하여 독음 순으로 정리하고 있다.

이형자는 이형의 정도가 경미한 글자가 여러 개있을 때에는 사용빈도가 높은 글자를 중심으로 정리한다. 이체자는 본체자가 있으나 역사적으로 혹은 현실적으로 필요에 의해 본체자의 형태를 변형하여 사용하는 글자를 가리킨다. 이형자는 이체자가 구조적인 차이가 있는 것에 비해 이형자는 필사자의 습관이나 역사적 변천에 따라 자형의 극히 일부 요소를 변형한 글자이다(이규옥. 2005).

통용자는 일반 통용자일지라도 색인용 테이블에서는 제외시키고 참고용으로만 제시하였다. 이체자가 글자 자형은 다르지만 정자와 동일한 음과 뜻을 가진 것이라면, 통용자는 글자 자형이 다를 뿐만 아니라 음과 뜻도 제한적인 상황에서만 동일하게 사용되는 글자이다. 일반통용자와 제한통용자로 구분할 때, 일반통용자는 이체자와 같이 쓰일 수 있는 것을 말하고, 제한통용자는 특정 문맥 속에서 혹은 다른 한자와 결합하여 어떤 단어를 형성할 때만 호환해서 사용할 수 있는 것을 뜻한다.

이체자사전 테이블은 그림 6에서 보여 지듯이 대표자와 이체자를 표시하고, 해당 이체자의 속성을 구분하여 나타내는 구조로 구성하였다. 검색엔진에 적용하여 검색을 수행하면 이체자가 모두 검색대상이 되는데 그 검색결과는 그림 7에서 볼 수 있다.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	연번	대표	부수	총획	이체1	구분	이체2	구분	이체3	구분	이체4	구분	이체5	구
1037	1036	從	彳	11	従	속	徙*	속	艾	동	从	본		
1038	1037	復	彳	12										
1039	1038	微	彳	13	徵	동								
1040	1039	德	彳	15	徳	속	悳	표	惠	속	恵	동	意	동
1041	1040	徵	彳	15	徵	동								
1042	1041	徹	彳	15	彻	간								
1043	1042	徼	彳	16										
1044	1043	忉	心	5										
1045	1044	忔	心	6										
1046	1045	忕	心	6	忬	동								
1047	1046	忌	心	7										
1048	1047	忘	心	7	忩									
1049	1048	怨	心	7	恣	고								
1050	1049	忍	心	7	忉	동								
1051	1050	忱	心	7	怗	속								
1052	1051	快	心	8	狭	본								

그림 6. 한자 이체자사전

그림 7 . 한자 이체자 사전을 적용한 검색결과

3.4 신출한자

한적 자료 구축 과정에서 유니코드로 입력이 불가한 문자를 신출한자로 규정하고 처리하고 있다. 신출한자를 수집, 검증과정을 거쳐 비표준 한자의 표현을 가능하도록 하여, 컨텐츠의 품질을 높이려는 노력의 일환으로 신출한자 처리를 지원하고 있다. 신출한자 중에는 우리의 문자생활에서만 필요한 한자를 만들어 사용한 것이 있는데, 이를 통칭 '고유한자(固有漢字)'라고 한다. 우리나라에서 고유한자가 만들어진 시기는 대체로 한자가 수입되었을 때부터라고 추정되고 있다. 이 과정에서 우리말의 음과 뜻을 보다 정확하게 표현하기 위하여 기존의 한자를 차용하거나 새롭게 한자를 造字하기도 하였다.

조자원리(造字原理)를 기준으로 구분하면, 크게 '형성(形聲)의 원리를 응용하여 만든(造字) 한자'와 '가차(假借)의 원리를 응용하여 만든 한자'로 나눌 수 있다. '형성의 원리를 응용하여 만든 한자'는 삼국시대부터 최근세에 이르기까지 오랜 역사에 걸쳐 형성되었는데, '[illegible](선)・[illegible](소)・畓(답)・[illegible](발)'과 같이 모양(形部)과 소리(聲部)의 조합형태로 만들어진 한자들을 가리킨다. '가차의 원리를 응용하여 만든 한자'는 다음의 네 가지 유형으로 나눌 수 있다(신상현, 2004).

① 우리말의 종성(終聲)을 동일한 자음의 한자로 가차하여 만든 한자 : 우리말을 이두(吏讀)로 표기할 때 많이 만들어졌으며, 고유한자의 주류를 이루고 있다.

終聲	漢字	造字例
～ㄴ	隱(卩)/分/叱	[illegible], [illegible], [illegible], [illegible] 등
～ㄹ	乙	乻, 乭, 乫, [illegible], [illegible], [illegible] 등
～ㅁ	音/味	[illegible], [illegible] 등
～ㅂ	邑(巴)/業/立	[illegible], [illegible], [illegible], [illegible] 등
～ㅅ	叱	[illegible], [illegible], [illegible], [illegible], [illegible], [illegible], [illegible] 등
～ㅇ	同	[illegible], [illegible] 등

※ '～ㅇ'은 '加應・末應・無應' 등과 같이 '應'자를 사용하여 표기하기도 하고, '～ㅁ'은 '古叱・今叱' 등과 같이 '叱'자를 사용하여 표기하기도 하였다.

② 한자와 한글 자모를 결합하여 만든 한자 : 훈민정음 창제 이후에 만들어진 고유한자로서, 초성・중성은 한자 1자로 나타내고 종성은 한글 자모를 사용하여 만들었다. 예를 들면 [illegible](걱)・[illegible](둑)・[illegible](억)・[illegible](둔)・[illegible](놈)・[illegible](웅) 등이 있다.

③ 기존의 한자를 차용하여 우리말의 음을 나타낸 한자 : 이두에서 우리 음을 표현하기

위해 차용한 것으로 '국음자(國音字)'라고도 한다. '釗(쇠)・串(곶)・卜(짐)・歙(삽)' 등과 같은 예와 향약명(鄕藥名)을 표기하기 위해 사용한 '召(棗)・干(薑)' 등과 같은 예가 있다.

④ 기존의 한자를 차용하여 우리말의 뜻을 나타낸 한자 : 기존의 한자에 우리나라 고유의 새로운 의미를 부여한 한자를 가리킨다. '국의자(國義字)'라고도 하며, '遷[벼랑]・評[고을]' 등이 있다.

역사자료 DB 구축 사업에서 새로 발견된 신출한자 중 고유한자로 판정되는 글자들의 유형은 다음과 같다.

첫째 우리말의 종성을 동일한 자음의 한자로 가차하여 만든 한자, 둘째 한자와 한글 자모를 결합하여 만든 한자, 셋째 특수하게 만들어진 인명한자의 용례가 발견되었다. 그리고 우리말의 종성을 동일한 자음의 한자로 가차하여 만든 한자는 다시 1) '隱(冂)/分/叱'을 '～ㄴ'으로 사용한 경우, 2) '乙'을 '～ㄹ'로 사용한 경우, 3) '音/味'을 '～ㅁ'으로 사용한 경우, 4) '邑(巴)/業/立'을 '～ㅂ'으로 사용한 경우, 5) '叱'을 '～ㅅ'으로 사용한 경우, 6) '同/應'을 '～ㅇ'으로 사용한 경우가 발견되었다.

우리나라 고유한자는 그 사용 범위가 인명, 지명, 물명 등 광범위할 뿐만 아니라 자형도 다양한 형태로 나타나고 있다. 이러한 고유한자의 유형들 중에서 가장 많은 비중을 차지한 인명용 한자는 다시 주로 하층계급의 이름을 표기하기 위해 한자의 음과 뜻을 가차한 것과 상층계급에서 특수하게 만들어진 것이 확인되었다.

음운학적 특징으로는 '叱'로 '～ㅅ'받침을 造字할 경우 '～ㄱ'받침으로 전이되는 경우도 나타나는데, '後 (뒷, 뒤)', '金 (쇳, 쇠)', ' 花(곶, 꽃)'의 예처럼 대체로 우리나라 고유음을 차용한 것이 많다. 그리고 '宋加里(송가파리)'의 '里 (파리)'처럼 특이하게 만들어진 경우도 있었다.

신출한자는 먼저 자료에 유니코드로 입력이 불가능한 한자가 출현하였을 경우, 일단 신출한자 후보로 등록하고 다음과 같은 처리 절차를 따른다(강대걸. 2005).

1) 원전 대조를 통한 오류 여부 검정
・원전 자형의 誤讀
・폰트 자형의 오류
・이체자 처리의 오류
・한자 정보의 오류
2) SuperCJK Ver.14.0와의 대조를 통한 신출한자 여부 판별
・해당 한자의 부수와 잔여획수를 확인하여 수록 여부를 통해 신출한자 여부를 판정
・획수 순서의 배열이 모두 일치하지 않으므로 ±1-2획까지 대조를 수행
3) IRG의 'Annex S' 문서 기준에 의거하여 이체자 통합 및 선별
・Annex S(IRGN951)는 유사 자형 한자와의 구분과 통합에 대한 기준

・각국에서 이체자를 무분별하게 제안하는 것을 통제하기 위한 것
・기존 이체자는 그대로 두고 이후 제안되는 신출한자에만 적용되는 기준
・통합해도 무방한 이체자와 분리해야 할 이체자를 일정한 규정을 만들어 표준화

4) Ext.C1 제출 한자 목록과 대조하여 Ext.C2 목록에 추가 등록
・Ext.C1 제출 한자 목록은 2002년 4월에 IRG에 제출한 우리나라의 신출한자 목록
・Ext.C1 제출 한자 목록과 대조하여 중복되지 않는 한자를 C2 제안한자 목록에 추가로 등록

5) 신출한자로 검출된 한자에 대한 문자정보 확인
・부수, 잔여획수, 총획수 등의 정보를 중심으로 재검토
・'국제 표준 제안한자 양식(IRG N881)'에 맞게 필요한 정보를 확인하여 입력

6) 신출한자 여부의 최종 판정 및 신출한자 목록 작성

2006년말 현재 5,980자의 신출한자를 확인하였고, 현재 표준코드가 없기 때문에 해당 폰트가 없으므로 문자의 이미지를 만들어 컨텐츠의 한자표현을 위해 이용하고 있다. 신출한자 검증결과 목록은 그림 8에서 볼 수 있다.

일련번호	임시코드	기관명	사업년도	DB명	출전 서명	출전 묶음명	위치 정보	파일명	전후문맥	원전 이미지	자형	파자	부수번호
06_규장각_0341	규장각_9067	규장각한국학연구원	2006	해설및해제자료	3권	규장각소장의궤해설해제	204398열	의궤_해제.xml	大膳에는 猪 1수, 牛後脚 1隻과 ★ 〈湔/乙〉 只, 鴨子 1수가 올랐으며	규장각_9067.hwp		湔/乙	005
08_전북대_0033	전북대_0033	전북대학교박물관	2006	고문서	1825年에 康津縣 列樹面에 사는 金翰豹가 作成한 戶口單子		jhm00237_a01	jhm00237.xml	賤口秩。婢日善年四十五辛丑。一所生奴雪/乙文年二十四	jhm00237_a01.jpg		雪/乙	005
10_한중연_0431	한중연_야사일기_0057	한국학중앙연구원	2006	조선조야사및일기자료		JE_A_24245_001	JE_A_24245_001_000061	a_kma_24245_001.xml	上士美院洞奮子歸★(庥/乙)<note type2="reader">麽</note>屯里</note>	JE_A_24245_001_000061.jpg		庥/乙	005
10_한중연_0154	한중연_궁중문화_0154	한국학중앙연구원	2006	한국궁중문화역사자료		JE_A_20118_008	JE_A_20118_008_005511	a_kma_20118_008.xml	等少壯男子願不及一老禿翁★(予+令)<note type2="reader">矜</note>眄閧鑠軍中頗	JE_A_20118_008_005511.jpg		予+令	006
08_전북대_0024	전북대_0024	전북대학교박물관	2006	고문서	1768年에 求禮縣 放光面에 사는 韓僣이 作成한 戶口單子		고14558_a01	14558.xml	放光面第一藍田里第一統統首申卞+口金	고14558_a01.jpg		亠/占	008
05_민추_0018	민추_0018	민족문화추진회	2006	한국문집총간	이재유고(頤齋遺藁)	한국문집총간 246집	mm_a246_008b-1	kc_mm_a557_av001.xml	庶此發軔而騁白▼(滅/木)兮。詔▼▼使指程。昔堯舜之嵬蕩兮。	mm_a246_008b-1		亠/儿/凵/冏	008
10_한중연_0095	한중연_궁중문화_0095	한국학중앙연구원	2006	한국궁중문화역사자료		JE_A_20046_001	JE_A_20046_001_000832	a_kma_20046_001.xml	沆<note>瑀子之</note>享王<note>臣下之享王自抗<note type2="reader">沆</note>始後世視以爲常以至上下★(亠/執/云)<note type2="reader">褻	JE_A_20046_001_000832.jpg		亠/執/云	008
02_국편_0003	kh7_003	국사편찬위원회	2006	한국근현대인물자료	대한민국인물연감 1편		310면 우1행	im_b012_008_0220	金龍⊙(亻+(氾-氵))	im_b012_008_0220.jpg		亻+(氾-氵)	009
05_민추_0206	민추_0206	민족문화추진회	2006	한국문집총간	청장관전서(靑莊館全書)	한국문집총간 258집	mm_a258_421a-1	kc_mm_a577_av051..xml	急以筆內闕爲彡。此藏拙法也。後人傳之。多書▼字。	mm_a258_421a-1		人+彡	009

그림 8. 신출한자 검증결과 목록

4. 결론

다국어를 컨텐츠를 대상으로 하는 정보시스템은 기존의 국내 표준 문자코드로는 문자를 자유롭게 표현할 수 없을 뿐만 아니라, 표준 문자코드를 이용하지 않고 자체 문자코드를 이용하는 경우에는 데이터 표현 및 교환이라는 측면에서 그 품질을 보장할 수가 없다. 이러한 문제를 극복하기 위해 많은 정보시스템이 유니코드 기반 문자코드를 적용하고 있거나 변환 작업 중에 있다. 이러한 과정에서 한글표현 및 한자표현을 위한 유니코드의 제한점으로 인해 이를 해결하기 위한 여러 방안이 모색되고 있다. 본 연구에서는 한자표현 및 검색을 지원하기 위한 방안으로 한자-한글음가 사전, 특수용어사전, 이체자사전, 그리고 신출자 목록 등을 제시하였다. 자료의 특성을 반영하는 이러한 검색 지원으로 정보표현과 더불어 정보검색의 품질을 확보하였다고 할 수 있다.

이러한 정보시스템의 문자표현과 관련한 문제를 해결하려는 노력에도 불구하고 현실적으로 다음과 같은 문제점이 나타나고 있다.

첫째, 표준 문자코드 적용과 관련한 문제로 현재 대부분의 정보시스템이 유니코드 기반으로 문자코드를 적용하거나 변환하는 과정 중에 있으나 일부 시스템은 여전히 기존의 문자코드를 적용하고 있다. 이들은 정보검색시에 이용자가 질의어로 대부분 한글만을 이용하는 특성이 있으므로 크게 상관없다는 논리로 대응하고 있다. 그렇다 하더라도 이는 대상 컨텐츠의 정확한 문자표현을 통한 고품질의 정보를 제공해야 하는 당연한 의무에 위배되는 결정이다. 대상 컨텐츠 중 극히 일부만이 한자 및 옛한글을 포함하더라도 정확한 문자표현을 통하여 컨텐츠의 품질을 보장함과 동시에 그 자료를 이용하는 이용자의 만족도를 높여야 할 것이다.

둘째, 호환 영역 한자 적용에 관련하여, 시스템마다 처리하는 방식의 차이로 인한 문제점이 발생한다. 흔글이 유니코드 기반 한자코드를 완벽하게 지원하고 있지 않은 상황에서 이용자가 F9 호환영역의 한자를 이용할 경우에 이를 어떤 방식으로든 변환과정을 거쳐 문제를 발생시키지 않도록 해야 한다. 여기에는 두 가지 방식이 있는데, F9 영역의 한자를 색인에서 제외하고 그 영역의 한자를 입력하면 표준 한자로 변환 후 검색을 수행하는 방법과 F9 영역의 한자를 색인에 포함하여 검색하게 하는 방법이다. 전자의 경우에는 이용자가 입력한 한자가 표준 한자로 변환되기 때문에 이용자가 입력한 한자 외에 표준 한자를 포함하여 검색결과로 제시되므로 재현율은 향상되나 정확률은 낮아질 수 있다. 해당 관련자 간의 이러한 차이점에 대한 논의를 통해 유니코드의 표준적인 적용 방안에 대한 협의가 필요하다. 호환 영역의 한자 적용 방식에 차이가 있는 시스템간의 분산검색이나 통합검색시 이에 대한 정확한 대응이 선행되지 않으면 검색시 전혀 기대하지 않던 결과를 가져올 수 있다.

셋째, 신출한자 검증과 관련한 문제로, 표준 문자코드에 해당하지 않는 한자가 출현하였을 경우에 신출자인지 검증을 거쳐서 적용하여야 하는데 전문적인 검증 기관 및 인력의 확보가 쉽지 않다. 이는 국가와 민간차원에서 공동으로 고품질의 컨텐츠 구축을 위한 지원이 필요한 부분이라 할 수 있다. 이는 표준 문자코드에 등록되어 있지 않은 한자를 이름으로 가지고 있는 개인이나 컨텐츠를 구축하려는 당사자 혹은 해당 주체가 스스로 해결할 수 있는 문제를 넘어서는 노력을 요구하는 문제이다. 국가차원에서 한자의 체계적인 정리와 검증을 위한 지원, 표준 문자코드에 등록과 시스템에 이를 적용하는 포괄적인 노력이 필요한 부분이다.

참고문헌

강대걸. 2005. 한국역사정보통합시스템 구축 및 포털시스템 업그레이드 사업 검색엔진 사전

테이블 업그레이드 연구. 국사편찬위원회.
기술표준원, “국제문자부호계 KS규격의 국제규격부합화 연구”, 한국표준협회, 2000.
류범종, 최윤수. 2004. 정보검색 관리시스템 KRISTAL-2002. 지식정보인프라 제 15호: 36-40.
문자코드연구센터. 2004. “신출한자의 처리와 국제 표준화”, 2004년 문자코드연구센터 워크숍.
박종우. 2007. “한자 국제 표준화와 인명용 한자“, 문자코드연구센터 소식 제 19호: 1-3.
산업표준심의회, “정보 교환용 부호계(한글 및 한자) KS X 1001”, 한국표준협회, 2004.
신상현. 2004. “우리나라 固有漢字 조사 보고(1)”, 문자코드연구센터 소식 제 14호: 7-8.
안대혁, 박영배. 2007. “유니코드 환경에서의 올바른 한글 정규화를 위한 수정 방안”, 정보과학회논문지, 소프트웨어 및 응용 제34권 제2호: 169-177.
이규옥. 2005. “한적자료 DB구축사업과 이체자 정리“, 문자코드연구센터 소식 제 15호: 2-4.
정우봉. 2005. 21세기 세종계획 문자코드 표준화 연구. 국립국어원.
조순영. 2003. 학술 데이터베이스의 유니코드 변환 적용에 관한 연구. 한국교육학술정보원.
주리정. 2001. “유니코드의 구조와 문제점”, 한국정보관리학회 제8회 학술대회 논문집, 23-28.
홍윤표, “한글코드에 관한 연구”, 국립국어연구원, 1995.
Allen J.D., Joe Baker, Richard Cook, Mark Davis, Michael Everson, Asmus Freytag, Jonh H. Jenkins, Mike Ksar, Rick McGowan, Lisa Moore, Eric Muller, Markus Scherer, Michel Suignard, and Ken Whistler. 2007 "The Unicode Standard 5.0," Unicode Consortium, Addison-Wesley.
Korpela, Jukka K. 2006. "Unicode Explained," O’Reilly.
Patrick Faltstrom and others, "Internationalizing Domain Names in Applications (IDNA) - RFC 3490," IETF, 2003.