공공정보시스템 부호체계 개선방안 연구

김지용*, 이송희**, 최진영***
*고려대학교 컴퓨터정보통신대학원 소프트웨어공학과
**고려대학교 컴퓨터・정보통신 연구소
***고려대학교 컴퓨터・통신공학부
e-mail:coolguy@korea.kr, {shlee,choi}@formal.korea.ac.kr

A Study on the Improvement of the Code System in Public Information Systems

Ji-Yong Kim*, Song-Hee Lee**, Jin-Young Choi***
*Dept of Software Engineering, Korea University
**Institute of Computer Information and Communications, Korea University
***Dept. of Computer Information & Communication, Korea University

요 약

공공정보시스템에서 외래어 표기법에 어긋나는 귀화자 성명이나 브랜드명(법인명)을 사용할 경우에, 비표준 확장한글을 인식하지 못하여 성명이나 주소를 포함하는 글자가 “?”로 표시되는 깨짐현상이 발생하여 공공서비스 이용에 많은 불편함을 초래하고 있다. 그 원인으로는 첫째로, 외래어 표기법을 준수하지 않았다는 것, 두 번째로는 기존에 구축된 공공정보시스템 대부분이 'EUC-KR' 인코딩 방식을 사용하고 있으므로 비표준 확장한글을 표현하지 못하고 있기 때문이다. 이에 본 논문에서는 기존의 운영환경을 그대로 유지하면서, 비표준 확장한글을 지원할 수 있는 시스템 운영방안을 제안하였다. 연구결과는 실제 공공정보시스템 운영시에 적용할 수 있으며, 사용자에게 보다 나은 서비스를 제공할 수 있다.

1. 서론

공공정보시스템 운영시 샾, 믜, 컼 등의 비표준 확장한글 사용으로 인해 회원가입에 필요한 실명확인, 온라인 발급서비스 등 서비스 이용에 제약사항이 나타나고 있다. 정보시스템 이용시 비표준 확장한글 사용이 이처럼 이슈화가 된 원인은 우리나라의 귀화자가 100만명에 넘어서고 있으며 외국인이 귀화하는 경우 대법원 ‘외국의 국호, 지명 및 인명의 표기에 관한 사무처리지침’ 제2조를 살펴보면 “가족관계등록부 및 가족관계등록신고서에 기록 또는 기재하는 외국의 국호, 지명 및 인명은 해당 외국의 원지음(현지발음)을 한글로 표기하되....“로 명시되어 있는데, 이처럼 귀화자 이름의 한글표기시 비표준 확장한글이 종종 사용이 되고 있다. 또한 브랜드 이미지의 인지도를 높이기 위해 차별화된 브랜드명 및 법인명을 사용하는 경우에도 비표준 확장한글을 종종 사용하고 있다. 따라서, EUC-KR 인코딩 방식을 사용하는 공공정보시스템에서 이러한 외래어 표기법에 어긋나는 귀화자 성명, 브랜드명, 법인명을 입력하게 되는 경우에는 시스템에서 입력값 그대로 표현하지 못하여 “?”로 표현되는 깨짐현상이 나타나고 있다. 이에 본 논문에서는 기존의 운영환경을 그대로 유지하면서, 비표준 확장한글 깨짐 현상을 해결할 수 있는 방안을 제시한다. 본 논문의 구성은 2장에서 공공정보 시스템에서의 비표준 확장한글 문제점에 대해 자세히 알아보고 3장에서는 해결방안을 제시하며 마지막 4장에서 결론을 맺는다.

2. 공공정보 시스템에서의 비표준 확장한글 문제점

공공정보시스템을 운영하는 과정에서 비표준 확장한글 표현시 깨짐 현상의 원인에 대해 자세히 알아보고, 이로인해 초래되는 문제점들에 대해서 알아보고자 한다.

2.1 외래어표기법 미준수

국립국어원 외래어표기법[1] 제3항 ‘표기의 원칙’을 살펴보면, “외래어 표기의 받침에는 ‘ㄱ, ㄴ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ’만을 쓴다.”라고 규정하고 있지만, 샾, 디스켵, 엔뚜엘한, 워컼스 등과 같이 귀화자의 성명, 브랜드 명, 법인명등 외래어 표기법에 어긋나는 표기를 사용하고 있어 정보시스템 이용시 문제가 발생하고 있다.

2.2 공공정보시스템 인코딩방식의 한계

가장 많이 사용되는 인코딩 방식은 EUC-KR, Code Page 949(CP949), UTF-8이다. EUC-KR(완성형 한글)은 KSC5601(KS X 1001)기반의 문자 인코딩방식으로, KSC5601이란 1987년 한국공업진흥청에서 국가표준으로 정한코드로 원래 코드명칭은 "KSC5601정보교환용부호(한글 및 한자)"이며 사용빈도가 높은 한글 2,350자를 골라 가나다순으로 배치하였고 한자 4,882자 및 특수문자 986자로 구성되어 있다. 이 체계는 제정이후 행정전산망 등에 사용되었다[2]. 그러나 EUC-KR 인코딩 방식은 그림 1과 같이 2,350자의 한글을 코드형태로 제공하기 때문에 빠진 글자가 많아 다양해진 현대한글을 표현할 수 없고 한글조직의

기본원리에 어긋나 있으며 등록된 글자를 화면에 나타내지 못하는 치명적 결함까지 갖고 있다[3]. 현재 우리나라에서 운영중인 국민생활과 밀접한 국가 주요 공공정보시스템(주민시스템, 자동차정보시스템, G4C 등)은 EUC-KR 인코딩 방식을 사용하고 있다.

BBD0	뿡	뿌	뿍	뿐	뿔	뿜	뿟	뿡	쀼	쀵	쁘	쁜	쁠	쁨	쁩	삐
BBE0	빽	뺀	뺄	뺌	뺍	뺏	뺑	사	삭	삯	산	살	살	삵	삶	삼
BBF0	삽	삿	샀	상	살	새	색	샌	샐	샘	샙	샛	샜	생	샤	
0x	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
BCA0		샥	샨	샬	샴	샵	샷	샹	섀	섄	섈	섐	섕	서	석	섞
BCB0	섟	선	섣	설	섦	섧	섬	섭	섯	섰	성	섶	세	섹	센	셀
BCC0	셈	셉	셋	셌	셍	셔	셕	션	설	셤	셥	섯	셨	성	셰	센
BCD0	셸	셍	소	속	솎	손	솔	솖	솜	솝	솟	송	솔	솨	쇽	솬
BCE0	솰	쇙	쇄	쇈	쇌	쇔	쇗	쇘	쇠	쇤	쇨	쇱	쇱	쇳	쇼	쇽
BCF0	숀	숄	숌	숍	숏	숑	수	숙	순	숟	술	숨	숩	숫	숭	
0x	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
BDA0		숯	술	숲	숴	쉈	쉐	쉑	쉔	쉘	쉠	쉥	쉬	쉭	쉰	쉴
BDB0	쉼	쉽	쉿	슁	슈	슉	슐	슘	슛	슝	스	슥	슨	슬	슭	슴
BDC0	슷	슸	슿	시	식	신	싣	실	싫	심	십	싯	싱	싶	싸	싹

(그림 1) KSC 5601 코드

CP949방식은 마이크로소프트사가 도입한 코드페이지이며 확장완성형 또는 통합형 한글코드라는 명칭으로 확대되어 현대의 모든 한글을 수용하지만, 이 코드 페이지는 IANA(Internet Assignede Numbers Authority)에 등록되지 않았다. 따라서 인터넷 상에서 정보를 주고 받는 표준이 아니며[5] 윈도우용으로 개발된 비표준 방식이다. UTF-8 방식은 입·출력시 변환이 불필요하고 자동 정렬도 간편하며 한글외에 다양한 문자도 동시에 표현가능하나 DB에 저장된 EUC-KR 데이터를 모두 UTF-8로 전환이 필요하며 전환시 용량이 증가(한글 1자당 2Byte → 3Byte)하여 추가 저장공간이 필요할 수 있다. 그러나 표현방식을 변경하는 방안은 모바일 등 다양한 환경에서 국제 표준으로 지원하는 UFT-8방식이 적합하다고 할 수 있다[6]. 표1과 표2는 각각 이들 세가지 인코딩 방식의 장단점과 현황에 대해서 정리하였다.

<표 1> 인코딩 방식의 비교

구분	EUC-KR	CP949	UTF8
장점	완성형 코드만을 입출력 하는 것으로 높은 성능 보장	2바이트로 모든 한글 저장입출력 가능하며 공간 소모가 적으면서 모든 한글을 입출력 가능함	한글 11,172자가 순서로 배열 가능하며 전세계 모든 언어 지원
단점	한글을 2,350자 밖에 지원하지 못함	완성형 호환으로 글자배열순서와 정렬 순서가 다름. "ORDER BY" 절로는 한글정렬 불가	저장 공간 소모가 큼

<표 2> 인코딩 현황

방 식	EUC-KR	CP949	UTF-8
비표준 확장한글	처리불가	처리가능	처리가능
시스템수	919개(62.1%)	317개(21.4%)	244개(16.5%)
표현글자수	한글 2,350자	한글 11,172자	한글 11,172자
한글 1자당용량	2 Byte	2 Byte	3 Byte

2.3 발생구간

비표준 확장한글 문제가 발생하는 구간은 메일 전송시 인코딩 타입에 따라 쓸 수 없는 문자가 발생하며, Cilent Application은 OS에 의존적으로 OS가 제공하는 환경에 제약되어 입·출력이 제한된다. 그리고 Database 설정 또는 버전에 따라 유니코드를 받지 못하는 경우도 있다[4].

2.4 시스템 운영환경 및 처리과정 소개

○ 운영환경

표 3에서 보는바와 같이 이용자 PC는 비표준 확장한글이 지원되는 Window(CP949)를 이용하여 처리되며 실제 처리가 되는 응용프로그램서버(AP서버), 연계서버 및 연계기관 시스템이 EUC-KR 인코딩 방식으로 구성되어 있다.

<표 3> 시스템별 운영환경

	이용자 PC	내부시스템	연계기관
인코딩방식	CP949	EUC-KR	EUC-KR

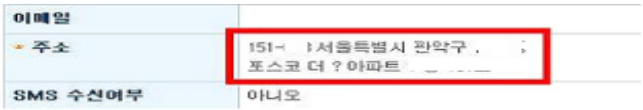
○ 처리과정

이용자가 회원가입으로 이용자 주소를 내부DB에 저장하거나 또는 온라인 발급서비스를 이용할 경우 연계기관을 통해 자료를 송수신 하는 처리과정을 표 4와같이 구성하였다. 내부 처리시 비표준 확장한글을 저장할 경우 AP서버가 '샲' → '?'으로 인지하여 회원DB에 저장된다. 또한 민원발급을 위해 연계기관 DB(원천데이터)에서 데이터 추출시에도 시스템간 연계과정을 거치기 때문에 발급문서에 출력되는 글자는 "샲"이 아닌 "?"가 되게 된다.

<표 4> 시스템별 처리과정

저장 시 (내부처리)	이용자PC	→	내부 AP서버	→	내부DB
	샲		?		?
외부처리 연계시	연계기관	→	내부 AP서버	→	발급문서
	샲 또는 ?		?		?

그림 2는 회원DB의 회원주소가 깨짐현상이 발생하여 저장되어 있음을 보여준다.



(그림 2) 회원DB정보

또한 시스템간 자료 연계시 한글 11,172자를 지원하는 인코딩방식인 CP949, UTF-8방식으로 저장된 자료라 하더라도 EUC-KR을 사용하는 시스템과 연계 처리를 할 경우, EUC-KR방식은 2,350자외의 한글을 지원하지 못하기 때문에 깨짐현상이 발생하여 시스템 연계시 문제가 발생하고 있다.

2.5 서비스 제약현황

정보시스템에서 비표준 확장한글 사용시 아래와 같이 서비스 제약사항이 발생한다.

○ 실명확인 문제(귀화인 성명, 법인명)

정보시스템의 원활한 서비스 이용과 익명사용자로 인한 피해를 방지하기 위해 회원가입 시 실명인증과정을 거치게 되는데 표 5와 같이 깨짐현상으로 인해 실명인증을 못하는 경우가 발생한다.

<표 5> 실명인증 처리과정

이용자 PC	↔	AP서버	↔	실명인증기관
(성명) 엔뚜엔한		엔뚜?한		(인증불가)

○ 민원서류발급 문제(특정 브랜드명)

민원서류 발급시에도 실명확인에서 발생하는 문제처럼 비표준 확장한글을 지원하지 못하여 표 6과 같이 발급되어 서비스 이용에 불편을 주게 된다.

<표 6> 민원서류 발급형태

현주소 : 서울특별시 관악구OOO 포스코 더? 아파트

3. 개선방안

클라이언트-서버(C/S), 웹(Web)등 단일 시스템에서 발생되는 비표준 확장한글 표현문제에 대해 시스템 특성을 고려하여 시스템 자체적으로 처리할 수 있는 방안마련과 정보시스템 간 한글코드 송수신 시 발생할 수 있는 비표준 확장한글 문제 해결방안을 그림3과 같이 마련하고자 한다[7].

방안	장점	단점	* 비고 - "포스코 더 샾 아파트" | UCS{ D3EC C2A4 CF54 0020 B354 0020 C0FE 0020 C544 D30C D2B8 }
KS X 1001 쪽자 방법 (채움+풀어쓰기)	KS 규격 만족하는 표준적 방법	입출력시 변환 필요.	"포스코 더 (채움)ㅅㅑㅍ 아파트" // (채움 : 0xA4D4)
HTML Character entity (NCR 표기법)	Web browser 운영시 효율적 대안	C/S 환경에서 변환 필요.	"포스코 더 샾 아파트" // U+C0FE 인코딩
Base64 Encoding 변환	글자 유지 가능	글자 데이터로서 의미를 잃음.	"포스코 더 lmN4Al 아파트" / / 0x98DE. 인코딩구분자 필요
CodePage 949 (윈도우 통합한글)	일반 한글윈도우 안에서 통용	윈도우 외 시스템 대비 안됨 타 시스템 연계시 깨짐	"포스코 더 샾 아파트" {C6 F7 BD BA C4 DA 20 B4 F5 20 98 DE 20 BE C6 C6 C4 C6 AE}
KS사용자 정의코드 확장	표준영역 코드 유지 시스템 변화 적음	코드 정의가 표준화 안되어 있고, 기 사용되고 있는 코드와 충돌 가능성. 입출력 재구성	"포스코 더 샾 아파트" // 0xC9A1사용자정의 첫번째 {C6 F7 BD BA C4 DA 20 B4 F5 20 C9 A1 20 BE C6 C6 C4 C6 AE}
UTF-8 인코딩 방법	모든 UCS글자 수용	기존 DB의 전면적 변환 필요	{ED 8F AC EC 8A A4 EC BD 94 20 EB 8D 94 20 EC 83 BE 20 EC 95 84 ED 8C 8C ED 8A B8 }

(그림 3) 확장한글 표현 및 연계방식 비교

기존에 사용하는 EUC-KR을 보완하는 방법에는 쪽자쓰기 방식과 NCR방식 등으로 처리할 수 있으며 쪽자방식과 NCR 표기방식은 기존 DB를 변경하지 않고 처리할 수 있는 장점을 가지고 있지만 쪽자쓰기 방식은 확장한글만을 표현하는 방법으로 많은 제약사항이 존재하고 입출력 변환시 변환모듈을 통해 처리가 가능하다. 이에 유니코드를 표현할 수 있고 웹 기반 시스템 수정이 용이한 NCR방식이 적합하다고 볼 수 있으나 한글 입·출력에 필요한 변환기능을 추가로 개발 적용하여야 하기 때문에 상당한 비용이 발생될 수 있고 DB에 확장한글이 숫자로 저장되어 있으므로 '가나다순' 자동정렬이 어렵다는 단점이 있다.[6]

이전에도 언급하였듯 국가 주요시스템의 60%이상이 EUC-KR을 사용하므로 본 논문에서는 시스템의 환경을 개선하기보다 운영환경을 그대로 유지하면서 비표준 확장한글을 처리하고자 한다. 해결방안으로 UTF-8 인코딩 디코딩 방식을 이용한 내부 및 외부처리(연계)에 대해 해결방안을 제시한다. 기존에 운영중인 시스템 환경(OS, MW, DBMS 등)은 표 7과 같다.

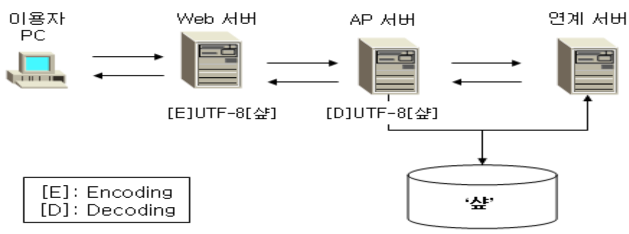
<표 7> 정보시스템 운영환경

	Web서버	AP서버	DB서버	연계서버
OS	EUC-KR	EUC-KR	EUC-KR	EUC-KR
미들웨어	EUC-KR	EUC-KR	-	EUC-KR
DB	-	-	CP949	-

DBMS의 인코딩방식이 CP949인 까닭은 시스템 OS 방식과는 별도의 DB 인코딩 방식을 구성할 수 있고 시스템 노후장비 교체에 따라 시스템장비 교체 및 DBMS 구성을 변경하여 기존에 사용하고 있는 DBMS를 오라클 8i에서 10g로 업그레이드 하였다. 그림3에서 제시한 방식은 시스템 환경이 모두 EUC-KR방식일 경우 해당되며 해결방안을 찾고자하는 본 시스템의 DB구성은 CP949 방식을 사용하기 때문에 비표준 확장한글 처리를 UTF-8방식으로 이용하고자 한다.

3.1 구현 시나리오

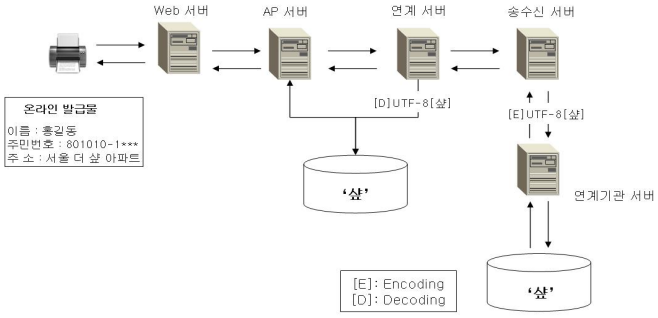
○ 내부처리(회원정보 저장) : 이용자는 Windows(CP949) OS를 사용한다는 전제하에 이용자의 회원정보가 그림 4와 같이 저장될 때 Web서버에서 AP서버로 데이터를 UTF-8 Encoding 방식을 이용하여 송신하게 되며 AP서버에서는 Ecoding으로 송신받은 데이터를 Decoding 하여 DB에 저장하는 방식을 이용한다.



(그림 4) 비표준 확장한글 내부처리 시나리오

○ 내부처리(실명확인) : 회원정보 저장과 같은 시나리오로 DB에 저장된 정보를 인증기관에 넘겨주고 결과값을 회신받게 된다.

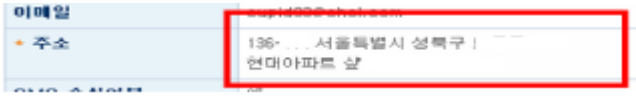
○ 외부처리 : 외부연계기관에서 추출된 자료에 대한 처리 과정은 그림 5와 같이 진행된다. 외부기관의 AP 서버에서 비표준 확장한글을 UTF-8 Encoding하여 연계서버를 통하여 자료를 수신 받게 되며 연계서버에서 Decoding하여 비표준 확장한글을 DB에 저장 후, 발급서식과 함께 출력한다.



(그림 5) 비표준 확장한글 외부처리 시나리오

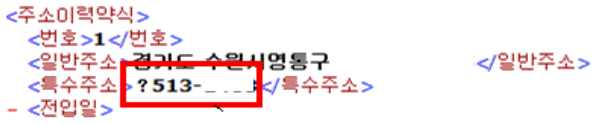
3.2 구현 및 처리결과

○ 내부처리 : 내부처리 시나리오를 통해 UTF-8 방식으로 인코딩·디코딩을 통하여 회원DB에 저장된 결과를 그림 6과 같이 확인할 수 있다.



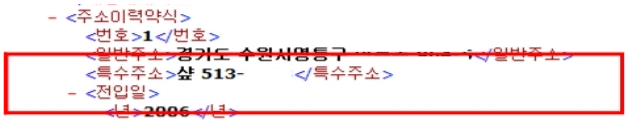
(그림 6) 적용 후 회원DB

○ 외부처리 : UTF-8방식의 적용 전인 XML데이터를 보여 주고 있다. 외부연계기관과 데이터 송수신시 인코딩 디코딩 방식을 사용하지 않아 그림 7과 같이 '샾' → '?' 처리되었음을 알 수 있다.



(그림 7) 외부처리결과 변경 전

그림 8은 UTF-8방식으로 외부연계 기관에서 인코딩한 후 연계시스템에서 디코딩하여 DB에 저장된 XML데이터를 보여주고 있다. 그림에서 보듯 '샾'→'샾'으로 처리되었음을 알 수 있다.



(그림 8) 외부처리결과 변경 후

3.3 서비스 개선시 고려사항

기존에 운영되고 있는 시스템이 UFT-8 인코딩 방식으로 변환이 되어 운영되는 것이 가장 좋은 방법이나, DB의 마이그레이션 비용과 클래스 파일의 재컴파일 작업을 필요로 하기 때문에 상당히 번거롭고 비용이 많이 소요된다. 이에 따라 기존의 운영환경을 유지하면서 데이터 처리 및 데이터 송수신시 국제표준 부호체계방식인 UFT-8을 이용하여 데이터를 인코딩·디코딩하여 해결하는 방안을 제시하였고 적용하는 과정에서 아래와 같은 이슈사항이 발생하였다. 첫째 원천데이터에서 비표준확장 한글이 깨져 저장되어 있는 경우 표준 데이터 송수신 방법을 이용하더라도 깨진 데이터를 복구할 수 없다. 둘째 상용프로그램(보안솔루션 및 발급솔루션 등)이 EUC-KR방식만 지원할 경우 표준방식으로 처리시 깨짐현상이 발생하였다. 이에따라 상용소프트웨어 벤더에 협조를 얻어 소스일부를 수정하여 UTF-8에서 처리가 가능하도록 지원을 받아야 한다. 그리고 외부처리(연계)시 원천데이터를 소유하고 있는 기관의 적극적인 협조(소스수정 및 테스트) 또한 중요하다.

4. 결론

정보시스템 구축운영지침('10. 5. 6)이 새로 개정되어 데이터 교환시 EUC-KR방식 뿐만 아니라 UTF-8방식을 제시하고 있으며, 신규 시스템 구축시 KS X ISO/IEC 10646(UTF-8) 사용 의무화 및 기존시스템의 UTF-8방식의 단계적 추진을 예정하고 있지만 국민생활과 밀접하게 연관되어 있는 국가 주요 정보시스템은 여전히 EUC-KR 방식으로 운영되고 있는 시스템이 많은 만큼 투자비와 작업량이 많고 연계시스템의 변경도 동반하게 된다. 따라서 EUC-KR 방식을 사용하는 시스템에서 비표준 확장한글을 오류없이 표현하고 송·수신할 수 있는 가이드라인이 작성·배포되어야 한다[7]. 또한 이미 축척되어 있는 데이터 정보의 마이그레이션 작업을 통해 시스템간 연계 처리시 서비스를 이용하는 이용자에게 불편함이 없도록 구체적인 전환계획이 수립되어야 하겠다.

참고문헌

[1] 국립국어원, "http://www.korean.go.kr"
[2] 한글부호체계, "http://in.rgall.com/itip-ch01/hcodes/hangulcode01.htm"
[3] 연합뉴스, "한글코드, 사용자위주로 바뀌어야", 1990
[4] NIA, "확장한글 문제 발생 개요-요약보고서", 2010
[5] 위키피디아, "http://ko.wikipedia.org/wiki/CP949"
[6] 행정안전부, "비표준확장한글 처리방안 내부회의자료", 2010
[7] NIA," 확장한글 개선방안 컨설팅 제안요청서", 2010