[인터넷] 다국어 도메인(IDN) 표준 개정 논의 동향

2002년 IETF의 다국어 도메인(IDN) 국제표준 확정 이후, 웹브라우저와 같은 응용프로그램의 다국어 도메인 비지원으로 인해 다국어 도메인 이용 활성화가 어려웠으나, 작년 다국어 도메인이 지원되는 인터넷 익스플로러 7.0이 출시됨에 따라 비영어권 국가의 다국어 도메인 이용이 크게 확산되고 있다.
ICANN(국제인터넷주소자원관리기관)에서도 비영어권 국가 인터넷 이용환경 개선을 위해 다국어최상위도메인(○○.한국 등)의 도입을 추진하고 있어 IETF도 이에 발맞추어 IDN 관련 국제표준의 제약 및 문제점들을 재검토하고 개정 작업을 진행 중이다. 본 고에서는 최근 진행 중인 다국어 도메인 관련 국제표준 개정 동향을 소개하고자 한다.

다국어 도메인(IDN) 개요 및 현 국제표준화 현황

다국어 도메인이란 도메인 이름의 영역에 영문이 아닌 한글과 같은 자국어를 사용한 도메인을 말한다. IETF에서는 2002년 영문 아스키(ASCII)만 인식 가능한 DNS 인프라스트럭처(infrastructure)의 구조적 변경이 어려움에 따라 응용프로그램에서 다국어 도메인을 영문자, 숫자, 하이픈(-)으로만 구성된 영문 퓨니코드(Punycode) 문자열로 변환 처리하는 방식을 선택하였다. 이같은 방식을 IDNA(Internationalizing Domain Names in Applications)라고 정하고 IETF 국제표준(RFC 3490)으로 채택하였다.
국내에서는 2003년 8월부터 한국인터넷진흥원에서 한글 도메인(한글.kr) 등록서비스를 제공중이며, 일반최상위도메인(.com, .net, .org 등)도 다국어로 도메인 등록이 가능하다.(한글.com 등)

현 국제표준의 문제점

현재 IDN 관련 RFC들은 2002년 당시에 발표된 유니코드 3.2에 의존적으로 작성되었다는 근본적인 문제점을 갖고 있다. 이미 유니코드 컨소시엄에서 새로운 문자셋이 추가되고 변경된 유니코드 5.0을 2006년 발표하였지만, 현 IDNA 표준은 이를 지원하지 못하고 있는 상황이다.
또한, 2005년 다국어 도메인은 "동형 이의어 스푸핑 공격(Homograph Spoofing Attacks)"에 대한 취약점이 발견되었다. 예를 들어, 유니코드 코드 포인트가 U+0430인 키릴 문자(Cyrillic) "a"는 유니코드 코드 포인트가 U+0061인 라틴어 소문자 "a"와 코드 포인트는 다르지만 모양이 동일하다. 이용자가 키릴 문자 다국어 도메인 "xn--pypal-4ve.com" 접속 시 웹브라우저 주소창에는 "paypal.com"로 표시되어 이용자를 혼동시키고 해킹에 악용될 소지가 있는 것이다.

이에 대해 모질라 계열 웹브라우저는 주소창에서 다국어 도메인을 영문 퓨니코드 형태로 표시하는 임시적인 해결방법을 사용하고 있다. (예: 한글.com 입력 시 xn--bj0bj06e.com 형

태로 변환하여 주소창에 표시) 그리고, 인터넷 익스플로러 7, 파이어폭스 2.0, 오페라 9.10 이상 버전에서는 주소가 의심스러운 경우 이용자에게 경고 메시지를 표시하게 된다.

국제표준 개정 추진 현황

기존 IDNA는 비허용 문자목록을 정해두고 다국어 도메인 입력 문자열 검사를 하였지만, 개정 IDNA에서는 허용 문자열 테이블을 만들고 이를 통해 입력 문자열의 처리여부를 검사하게 된다. 서로 혼동을 일으키지 않는 문자열만 허용하기 위해서이다.

문자열 테이블은 허용(always), 허용불가(never), 미정/추후 허용가능(maybe yes), 미정/추후 비허용 가능(maybe not) 4가지 종류로 분류되며, 기호, 구두점, 박스문자 등의 비언어 문자들을 제외한 오직 언어 문자만이 허용 대상이다. 이 테이블은 유니코드가 버전업 됨에 따라 필요 시 IDNA 입력 허용 문자열 목록은 계속 확장될 것이다.

다국어 도메인의 처리를 위해서는 미리 정의된 규칙에 따라 입력된 다국어 문자열의 공백 제거, 구두점 삭제, 대소문자 변환 등을 처리하는 텍스트 정규화(Text Normalization) 과정이 필요하다. 개정 IDNA는 문자를 호환성 등가(Compatibility Equivalence)로 분해 후, 규범적 등가(Canonical Equivalence)로 재구성하는 NFKC(Normalization Form Compatibility Composition) 방식 대신 문자를 규범적 등가로 분해/재구성하는 NFC(Normalization Form Canonical Composition) 방식을 사용하게 된다. (시각적으로 구별이 불가능하고 텍스트 비교 및 렌더링의 목적상 정확히 동일한 의미를 가지는 문자를 규범적 등가라고 하고, 동일한 문자나 문자 시퀀스의 대체 표현 문자를 호환성 등가라고 한다.)

규범적 등가의 예로는 'A'에 옹스트롬(Angstrom) 사인이 있는 'Å'가 있는데, 'Å'(유니코드 U+212B)와 라틴어 'Å'(유니코드 U+00C5)는 모양이 같고 코드 포인트가 다르지만 분해된 결과는 'A'와 ''(U+030A)로 같으므로 규범적 등가이다. 호환성 등가의 예로는 숫자 '2'(유니코드 U+0032)와 상첨자 '²'(유니코드 U+00B2)가 있는데, '²' 는 숫자 '2'의 다른 형태이긴 하지만 시각적으로 구별되고 의미도 다르기 때문에 규범적 등가에 해당되지 않는다. '2²'를 NFKC로 정규화하면 '²'가 호환성 등가인 '2'로 분해된 후 결합되어 '22'로 바뀔 수 있으나, NFC 방식 정규화는 규범적 등가를 사용하기 때문에 '2²'가 그대로 유지된다. 라틴어 단일 합자인 'ﬁ'의 경우에도 NFKC에서는 'f'와 'i'의 결합으로 표현될 수 있으나, NFC에서는 그렇지 못하다.

위와 같은 정규화 방식 변경으로 인해 기존 IDNA에서 등록이 허용되었던 일부 다국어 문자열이 개정 IDNA에서는 허용되지 않는 경우가 발생할 우려가 있다.

향후 표준화 추진계획 및 결언

유니코드 컨소시엄에서는 새로운 언어를 표현하기 위한 문자셋을 계속 유니코드에 추가할 예정이다. 2003년도 유니코드 3.2에 의존적인 IDN 관련 국제표준들은 현 유니코드 5.0뿐만 아니라 향후 유니코드 버전까지도 고려하며 계속적인 논의가 필요하다. 즉, 유니코드 버전에 관계없이 적용 가능한 IDNA의 개발이 필요할 때이다. 이를 위해 IETF에서는 현재 "xn--" 형태의 다국어 도메인 구별 접두사(prefix)를 다른 접두사로 변경하여 새로운 유니코드 버전을 수용하는 것까지도 고려하고 있다.

2007년 9월 ITU-T Study Group 17에서는 각국 정부 다국어 도메인 전문가들이 모여 다국어 도메인 관련 표준 개정안의 이슈 사항을 논의할 예정이며, 차기 IETF 회의는 2007년 12월 캐나다 벤쿠버에서 개최되고 IDNA 개정안에 대한 비영어권 국가의 의견수렴 및 허용 문자열 목록에 대한 논의가 계속 될 예정이다.

다국어 도메인의 응용프로그램 레벨에서의 처리를 다루는 IDNA RFC 개정은 향후 다국어최상위도메인 및 다국어 전자우편주소 국제표준화에까지 광범위한 영향을 미치게 되므로, 한글 도메인 주소의 올바른 표현 및 확장을 위해 IETF, 유니코드 컨소시엄과의 긴밀한 협조와 논의가 중요시 되는 시점이다.

김재연 (한국인터넷진흥원 연구원, ITU-T SG17 Associate Rapporteur, kimjy@nida.or.kr)