Kiedy e-mail do grażyna@księgarnia.łódź?

Nadchodzi duża zmiana, której my – użytkownicy języka polskiego – nie zapoczątkowaliśmy, ale z której możemy – na równych z innymi prawach – skorzystać. Doświadczony surfer wie, skąd i kiedy nadchodzi dobra fala. Ona już się zbliża – to czas, by zacząć surfować po internecie także w swoim stylu, korzystając w pełni ze znaków alfabetu języka polskiego.

Internet narodził się jako anglojęzyczny. Jednak dla szybko rosnącej populacji jego użytkowników angielski nie jest językiem ojczystym. Cyfryzacja obejmuje coraz to nowe dziedziny i coraz trudniej funkcjonować bez internetu i „adresu z małpą" @, który pozostaje przepustką do cyfrowego świata. Niwelowanie wykluczenia cyfrowego polega także na otwarciu tego świata na miliardy użytkowników języków innych niż angielski i alfabetów niełacińskich.

Pierwszy miliard użytkowników internet osiągnął w 2005 r. Dziś w samych Chinach dostęp do sieci ma miliard ludzi. Przez dziesięć ostatnich lat (2013–2023) liczba użytkowników globalnej sieci informacyjnej wzrosła ponad dwukrotnie – z 2 563 mln do 5 400 mln[1].

Ten dynamiczny wzrost był głównym, obok przyrostu wolumenu dostępnych treści, czynnikiem wymuszającym zmiany infrastruktury internetu. Limitowana przestrzeń adresowa protokołu IPv4 (przyznawanie adresów z tej puli regionalnym rejestratorom zakończyło się w 2011 r.) oraz ograniczenia pierwotnej, ściśle hierarchicznej struktury nazw domen zmusiły ICANN, organizację zarządzającą ładem w systemie DNS i przyznawaniem nazw domen internetowych, do szukania nowych rozwiązań. Już w latach 90. XX w. opracowano, a od 2011 r. rozpoczęto wdrażanie protokołu IPv6 wykładniczo zwiększającego liczbę adresów IP.

Mikołaj Karłowski

autor, zawodowo związany z instytucjami europejskimi, nie jest specjalistą IT, lecz użytkownikiem internetu, usług IT oraz języka polskiego. Pragnie być pełnoprawnym cyfrowym obywatelem. Tekst prezentuje poglądy autora.

Od 2000 r. ICANN rozpoczęła wprowadzanie do użytku nowych nazw domen funkcjonalnych (generycznych, gTLD) najwyższego poziomu. Proces ten przyspieszył w 2013 r. i obecnie jest ich ponad 1200. Charakteryzują się m.in. tym, że ich nazwy mogą być dłuższe niż dotychczasowe trzy znaki. Tak więc obok siedmiu „starych" domen: .com, .org, .gov., .edu, .mil, .net, .int internauci mają do dyspozycji setki nowych, takich jak np. .info, .online, .photography, .tokyo czy .brussels. Istnieją również domeny generyczne składające się ze znaków alfabetów niełacińskich, w tym np. arabskiego, pisanego od prawej do lewej strony. Nazwy domen najwyższego poziomu nie muszą już zawierać wyłącznie znaków ze standardu ASCII, ograniczonego do 26 liter alfabetu angielskiego.

1 Por. https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Równolegle rozpoczęto prace nad aktualizacją istniejącego od lat 90. XX w. protokołu dotyczącego międzynarodowych nazw domen (IDN), dopuszczającego znaki spoza standardu ASCII. Zaowocowało to utworzeniem w 2010 r. pierwszych międzynarodowych nazw domen krajowych najwyższego poziomu (IDN ccTLD). Obecnie istnieje ponad 60 takich domen krajowych. Wszystkie zapisane w alfabetach niełacińskich, głównie w arabskim i chińskim.

Unicode – w kierunku uniwersalizacji

Języki i pismo są bogactwem naszej cywilizacji. Jednak u zarania informatyzacji skąpość zasobów, w szczególności dostępnej pamięci oraz koszty transmisji danych wymuszały szereg ustępstw i odcisnęły piętno na sposobie, w jaki tworzono systemy, oprogramowanie, kodowano i przechowywano informację. To głównie z powodu ograniczeń technicznych pierwszych systemów, ale i dla wygody anglojęzycznych (amerykańskich) twórców narzędzi informatycznych i teleinformatycznych, stworzono i upowszechniono 7-bitowy standard kodowania znaków i wymiany informacji ASCII.

Wprawdzie na początku używanie 128 znaków ASCII (z czego 95 było znakami drukowalnymi) wystarczało Amerykanom, ale już Kanadyjczycy oraz użytkownicy europejscy, używający znaków diakrytycznych, rozszerzyli ten standard, tworząc zestawy 8-bitowe mogące zawierać do 256 znaków. To rozwiązanie powodowało niekompatybilność standardów i problemy w wymianie informacji. Ostatecznie usunęło je dopiero stworzenie standardu Unicode, obejmującego większość stosowanych na świecie znaków pisarskich, w tym logogramów, oraz popularyzacja przez Konsorcjum Unicode standardu kodowania znaków UTF-8, zaprezentowanego w 1993 r. UTF-8 jest od 2008 r. najpopularniejszym systemem kodowania znaków w zastosowaniach WWW, używanym przez ponad 98 proc. stron internetowych, dominującym także w alfabetach niełacińskich. Przypomnę, że bardzo ambitny i niezmiernie użyteczny standard Unicode, który może pomieścić aż milion znaków, zawiera obecnie bez mała 150 tys. znaków spośród 161 systemów pisma.

> *Unicode ujednolicił (zuniwersalizował), a tym samym uprościł wymianę informacji między użytkownikami różnych języków, żadnego z nich nie pozbawiając znaczenia i specyfiki.*

Universal Acceptance

Eksperci ICANN dostrzegli, że pomimo utworzenia nowej puli zasobów ich recepcja i wykorzystanie, a tym samym rozwój internetu, napotykają przeszkody, w tym bariery techniczne. Dotyczyło to zarówno domen generycznych, jak i międzynarodowych.

Obserwacje te sformułował w trzech punktach Ram Mohan, późniejszy założyciel Universal Acceptance Steering Group:

1. **stara domena najwyższego poziomu (TLD) będzie akceptowana przez systemy częściej niż domeny nowe;**
2. **nazwa TLD zawierająca jedynie znaki ASCII będzie akceptowana częściej niż domena międzynarodowa (IDN TLD);**
3. **stare domeny dwu- lub trzyznakowe będą akceptowane częściej niż dłuższe nazwy domen krajowych (ccTLD) lub domeny generyczne (gTLD).**

Według Mohana tak długo jak powyższe twierdzenia będą prawdziwe, cele Universal Acceptance (UA) nie będą osiągnięte.

Ambicją Mohana i współpracowników było, by każda domena najwyższego poziomu funkcjonowała poprawnie we wszystkich aplikacjach, niezależnie od użytego alfabetu, liczby znaków czy daty utworzenia. Tak określono główny cel Universal Acceptance Steering Group (https://uasg.tech) utworzonej w 2015 r. przy ICANN i skupiającej przedstawicieli firm cyfrowych, zainteresowanych rządów i społeczności.

UA ma już niemal dziesięcioletni dorobek, który jednak powszechnie nie jest znany, czy to z powodu specyfiki podejmowanych zagadnień, czy to z uwagi na koncentrację aktywności na regionach innych niż Europa i alfabetach innych niż łacińskie. Na ile upowszechniły się zasady Universal Acceptance? Można zaryzykować stwierdzenie, że z domenami problem właściwie jest rozwiązany. System DNS, dzięki algorytmowi Punycode, pozwalającemu na „tłumaczenie" nazw z Unicode na standard ASCII, obsługuje zarówno nazwy domenowe w ASCII, jak i w innych pismach. Dostępne są nowe TLD, także w innych alfabetach. Nie ma więc przeszkód formalnych ani technicznych, które wzbraniałyby np. rejestracji domeny TLD .łódź.

Internacjonalizacja adresów

Trzy *prawa Mohana* odnoszą się nie tylko do domen, lecz również do adresów e-mail, co czyni problem jeszcze bardziej złożonym. Jednym z celów UASG jest zachęcenie dostawców usług poczty elektronicznej, by umożliwili użytkownikom korzystanie z adresów zawierających domeny międzynarodowe i znaki spoza ASCII. Proces ten znany jest pod nazwą E-mail Address Internationalisation (EAI).

Działania skupiają się głównie na alfabetach niełacińskich – tam jest największe pole do zmiany – jednak przy okazji po-

stęp może się dokonać i z pożytkiem dla korzystających z pisma łacińskiego we wszystkich jego odmianach językowych.

> *Według UASG z technicznego punktu widzenia chodzi o przyjmowanie, zatwierdzanie, przetwarzanie, przechowywanie i wyświetlanie wszystkich nazw domen w sposób niezmienny i prawidłowy. Sformułowali więc pentadę:* ***acceptance, validation, processing, storing and displaying.***

Nie jest to jednak proste: dostosować trzeba w tym celu wiele elementów w bardzo różnych miejscach wielu systemów, począwszy od serwerów pocztowych, programów obsługujących pocztę u użytkowników końcowych, oprogramowania pocztowych serwisów webowych, praktycznie wszystkich pozostałych programów i usług, w których użytkownicy podają adresy e-mail.

Problematyka podejmowana przez Universal Acceptance ma doniosłe znaczenie w sytuacji, gdy adresy poczty e-mail stały się przepustką do usług świata cyfrowego i są powszechnie wykorzystywane do uwierzytelniania użytkowników. Szacuje się, że obecnie na świecie funkcjonuje ok. 8 miliardów kont e-mail. Oznacza to, że ponad połowa światowej populacji ma do czynienia z pocztą elektroniczną. By dopełnić obrazu: dziennie wysyłanych jest niemal 350 miliardów wiadomości e-mail, z czego niestety prawie połowa to spam. Cel postawiony przez UASG dotyczy więc praktycznie wszystkich internautów, korzyścią jednak będzie zmniejszenie wykluczenia cyfrowego głównie wśród użytkowników spoza obszaru języka angielskiego.

Adres do Grażyny

Problem walidacji adresów jest szczególnie istotny z perspektywy użytkownika. Przenieśmy zagadnienie na grunt polski: nadal w przeważającej większości przypadków nie jest możliwe tworzenie ani używanie nazw skrzynek pocztowych składających się z liter spoza zbioru ASCII. Jesteśmy zmuszeni korzystać z nazw skrzynek takich jak „sprzedaz" zamiast „sprzedaż". Nadal nie możemy rozróżnić, czy w adresie chodzi o sprzedaż pączków czy może odbiór paczek („paczki"). Żaden Mikołaj nie może wysłać e-maila do żadnej Grażyny. Adres e-mail taki jak grażyna@księgarnia.łódź, chociaż najzupełniej poprawny, nie może funkcjonować nie tylko z tego powodu, że jeszcze nie ma domeny TLD .łódź, lecz głównie dlatego, że operujący w Polsce dostawcy usług pocztowych nie obsługują, a przede wszystkim nie pozwalają na tworzenie nazw skrzynek zawierających polskie znaki diakrytyczne, mimo że właściwe standardy zostały przyjęte już z górą dziesięć lat temu.

Krótki przegląd zidentyfikowanych przez UASG problemów, jakie programiści i środowisko IT napotykają i muszą brać pod uwagę przy wdrażaniu UA do poczty elektronicznej[2]:

Obsługa protokołu SMTPUTF-8, będącego rozszerzeniem protokołu SMTP i umożliwiającego obsługę międzynarodowych adresów e-mail oraz nagłówków e-mail w formacie UTF-8. Choć protokół ten został wprowadzony w 2012 r., to jednak nadal stosuje go niewielu dostawców usług e-mail.

Unicode Bilateral Algorithm – zasady pozwalające na rozwiązywanie konfliktów związanych z zapisem od strony lewej do prawej (np. w jęz. arabskim czy hebrajskim), od prawej do lewej oraz dwukierunkowym.

Stosowanie algorytmu Punycode, umożliwiającego przepisywanie nazw domen zapisanych w Unicode na ASCII. Nazwy domen zapisywane są w rekordach DNS odpowiednio jako A-Labels (ASCII) i U-Labels (Unicode). A-Labels zawierające nazwy międzynarodowe zaczynają się od przedrostka „xn--". Stworzono biblioteki programistyczne ułatwiające wykorzystanie tego algorytmu.

Serwery pocztowe (Mail Transfer Agents). Proponowane jest w pierwszej kolejności ogłaszanie przez serwery pocztowe obsługi protokołu SMTPUTF-8. Gdy protokół nie jest obsługiwany, serwer wysyłający powinien wysłać wiadomość w „starym" formacie, tj. z adresem zawierającym jedynie znaki ASCII i bez nagłówków zawierających kodowanie UTF-8.

Programy pocztowe (Mail User Agents). Tu największym wyzwaniem jest obsługa nazwy skrzynki pocztowej (łańcucha poprzedzającego @, tzw. local-part w adresie e-mail) w Unicode.

Warto też wspomnieć o kwestii prawidłowej walidacji i przetwarzania łańcuchów będących nazwami domen i adresami e-mail w linki w treści wiadomości e-mail i innych przypadkach użycia (tzw. linkification).

2 W tej części artykułu korzystałem z prezentacji *UASG 019B Email Address Internationalization – Technical Perspective EN* https://uasg.tech/download/uasg-019b-email-address-internationalization-technical-perspective-en/

W 2014 r. NASK zarejestrował ok. 50 000 nazw domen ze znakami diakrytycznymi, obecnie aktywna jest połowa tej liczby[3]. A to i tak ułamek wszystkich aktywnych nazw w domenie .pl; jest ich ok. 2,5 mln. Dlaczego tak mało polskich domen IDN zarejestrowano? Użytkownicy nie bardzo wiedzieli, jak z takich domen korzystać, skoro nie akceptowały lub nie wyświetlały ich poprawnie przeglądarki internetowe i inne programy. Obecnie ten problem został częściowo rozwiązany – przeglądarki coraz lepiej radzą sobie z domenami międzynarodowymi. Jednak ich wykorzystanie pozostaje znikome z innych powodów: nadal zbyt często takich nazw nie akceptuje oprogramowanie różnych serwisów i systemów; ograniczenia nadal stosują dostawcy usług i oprogramowania trzymający się starych standardów. Dotyczy to również sytuacji, gdzie wykorzystywane są adresy poczty elektronicznej. Być może dlatego polscy użytkownicy nie widzą wielkiego sensu wykorzystywania międzynarodowych nazw domen.

Przypomnijmy, że polski rejestr NASK był jednym z pierwszych na świecie i pierwszym w Europie, który dopuścił w 2003 r. w domenie .pl nazwy międzynarodowe ze znakami diakrytycznymi w alfabetach m.in. polskim, greckim, hebrajskim oraz w cyrylicy. Szkoda, że obecnie nie przewiduje tworzenia domen drugiego poziomu z polskimi diakrytykami. W ten sposób wykorzystanie polskich znaków byłoby pełniejsze.

Tworzenie domen z polskimi znakami oferuje również rejestrator europejski (.eu). Niestety, przeważająca większość rejestratorów domen funkcjonalnych nie umożliwia (jeszcze?) zakładania takich domen.

Obecnie najczęściej spotykanym rozwiązaniem wykorzystującym IDN jest aliasowanie nazw domen, czyli przekierowanie zapytań użytkowników z domeny ze znakami diakrytycznymi do domeny głównej. To niewiele, ale lepiej niż nic. Polskojęzyczni użytkownicy przyzwyczaili się do kaleczenia polszczyzny i wymawiania oraz zapisywania tych nazw (oraz adresów) bez diakrytyków.

Czy problem z perspektywy użytkowników języka polskiego jest marginalny? Być może, ale chcę zwrócić uwagę, że na świecie rozpowszechnią się międzynarodowe adresy e-mail zapisane w innych systemach pisma. Czy będą prawidłowo obsługiwane w Polsce? Może należy dostosować działające w Polsce systemy w taki sposób, by nie zderzyły się z nieuniknioną zmianą, a przede wszystkim – by mieć pewność, że we właściwy sposób zagwarantujemy interes użytkowników języka polskiego?

W ruchu Universal Acceptance chodzi w istocie o to, by usunąć bariery rozwojowe (głównie techniczne) stojące na drodze upowszechnienia internetu i umożliwienia dostępu do niego kolejnym rzeszom użytkowników oraz by podnieść satysfakcję z usług dotychczasowych użytkowników. UASG mówi skromnie o kolejnym miliardzie nowych internautów, którzy mogliby korzystać z sieci we własnych językach. A dzisiaj internauta to właściwie to samo co obywatel. Tak więc udostępnienie ludziom internetu w ich własnym piśmie uczyni z nich pełnoprawnych cyfrowych obywateli.

Użytkownikom alfabetów niełacińskich trudniej sprostać wyzwaniom Universal Acceptance. Może dlatego wyraźniej dostrzegają oni problem i korzyści płynące z jego rozwiązania? Większość aktywności związanych z Universal Acceptance Day (przypada 28 marca) ma miejsce w krajach Azji, Afryki i Ameryk. W Europie, sądząc po zorganizowanych do połowy 2024 r. spotkaniach[4], zainteresowanie przejawili Francuzi (UNESCO), Hiszpanie, Macedończycy, Serbowie i Szwajcarzy. Europejscy, w tym polscy, zarządzający ładem w internecie nie powinni jednak lekceważyć tego zagadnienia.

Projektowi Universal Acceptance można by zarzucić niszowość. Nie dotyczy bezpośrednio naszego obszaru, piętrzy trudności, grozi wzniesieniem nowej wieży Babel, która spowoduje, że ludzie nie będą w stanie pokonać barier odmiennych pism i języków, komplikuje życie, ma charakter antyuniwersalistyczny.

Sądzę jednak, że taka krytyka nie jest uprawniona. Podobnemu przedsięwzięciu, Unicode, które odniosło niekłamany sukces, trudno przecież zarzucić antyuniwersalistyczny charakter. Przeciwnie: waloryzuje on, przechowuje i rozpowszechnia dorobek cywilizacyjny całej ludzkości, jakim są rozliczne systemy pisma, dodatkowo robiąc to w sposób niedyskryminacyjny. Myślę, że to jest ścieżka, którą powinniśmy podążać. Czas więc na dopuszczenie do powszechnego użytku adresów e-mail ze znakami narodowymi i powszechniejsze użycie międzynarodowych nazw domen. Konieczne byłoby np. umożliwienie użytkownikom poczty elektronicznej aliasowania ich skrzynek pocztowych: korzystania zarówno z międzynarodowego (Unicode), jak i dotychczasowego adresu na podstawie opracowanych, także przez polskich ekspertów, zasad i polityk, którymi mogliby kierować się operatorzy systemów. Ostatecznie aliasy to nic rewolucyjnego.

Przyszła pora na działania polskich dostawców usług pocztowych, twórców oprogramowania i aplikacji, a także na zachętę ze strony liderów środowiska IT i użytkowników. Czas najwyższy, by Grażyna mogła korzystać z porządnego adresu poczty elektronicznej.

3 Por. https://idn.pl/statystyki/liczba_aktywnych_IDN

4 Por. https://uasg.tech/ua-day/