Internet w bibliotekach II
łączność, współpraca, digitalizacja,
Wrocław, 23-26 września 2003 roku

Marek Nahotko

Energoprojekt - Kraków S.A.

Identyfikacja obiektów w sieciach rozległych

Identyfikatory są nadzwyczaj przydatnym narzędziem ułatwiającym komunikację pomiędzy użytkownikami Internetu. Nie są one jednak oryginalnym wytworem technologii internetowych. Wszystkim znany jest International Standard Book Number (ISBN), który stanowi główny pomost ułatwiający komunikację pomiędzy księgarzami, wydawcami i bibliotekarzami. International Standard Serial Number (ISSN) również spełnia podobną rolę w wypadku wydawnictw ciągłych; jest on również niezbędny w realizacji wewnętrznych procesów bibliotecznych, takich jak akcesja czasopism. Numery identyfikacyjne wykorzystywane dla celów bibliografii, takie jak numery OCLC czy RLIN służą do wykrywania zdublowanych opisów i ich konsolidowania podczas tworzenia bazy danych katalogów centralnych online[1].

Jednoznaczne i trwale niezmienne (ang. *persistent*) identyfikatory dla obiektów cyfrowych są zasadniczą częścią technologii informacyjnej[2]. Pozwalają one na:

- handel elektroniczny, w tym transakcje prawne;
- zarządzanie prawem autorskim (copyright);
- sporządzanie elektronicznych spisów treści;
- zarządzanie produkcją i inne działania administracyjne wewnątrz instytucji;
- działalność bibliograficzną i przeszukiwanie zasobów.

W środowisku sieciowym niezbędne są co najmniej następujące kategorie identyfikatorów:

- Identyfikatory autorów. International Standard Authority Data Number (ISADN) unikalnie identyfikuje każdego autora. Jest to bardzo istotne w wypadku istnienia kilku, równie "legalnych" form nazwy autora, np. z powodu transliteracji. ISADN znacznie ułatwi prowadzenie wyszukiwania w katalogach centralnych wg nazwiska autora, jest on również niezbędny dla właściwego pobierania opłat za prawa autorskie.
- Identyfikatory dzieła. Te identyfikatory niezbędne są z powodu istnienia dużej liczby różnych materializacji tego samego dzieła. W ISO trwają obecnie prace nad International Standard Work Codes. Mówi się o trzech standardach z tej grupy: International Standard Audiovisual Number (ISAN), International Standard Musical Work Code (ISMC) i International Standard Textual Work Code (ISTC). Systemy obsługujące te identyfikatory powinny także spełniać wymogi FRBR.
- Identyfikatory materializacji prac, takich jak m.in. wydania książek. W tej klasie identyfikatorów znajdują się ISBN i ISSN.
- Identyfikatory dla części składowych materializacji. W tym zakresie istnieją dwa identyfikatory: SICI i BICI[3].

Zarządzanie całym zestawem identyfikatorów przedstawionych powyżej stanowi olbrzymią pracę. Dla właściwego funkcjonowania wszystkie systemy identyfikatorów muszą być zaopatrzone w metadane. Na przykład ISSN nie posiada znaczenia bez metadanych opisujących wydawnictwo seryjne, dla którego przydzielono identyfikator. ISSN wymaga, aby każde wydawnictwo posiadające ten identyfikator zostało skatalogowane. Na razie nie ma takiego wymogu dla ISBN-u, jednak nowy projekt ISBN-u przewiduje obowiązek dołączania metadanych ONIX (Online Information eXchange)[4].

W literaturze z problemami identyfikacji związanych jest zazwyczaj kilka zagadnień. Są to:

- Niezbędny poziom szczegółowości. W systemie tradycyjnym wydawcy pracowali na poziomie książki lub czasopisma, używając jako identyfikatorów ISBN lub ISSN. Jednak publikacje (opisywane jednostki) mają coraz mniejszą objętość. Obecnie możliwa jest identyfikacja pojedynczych artykułów. Pojawia się potrzeba identyfikacji znacznie mniejszych fragmentów prac, np. części tekstu czy oprogramowania, a także pojedynczych ilustracji, obrazów, wideoklipów itp.
- Identyfikatory są "głupie" lub "inteligentne". "Głupi" identyfikator nie posiada właściwego sobie znaczenia, więc może być tylko użyty do wyszukania go w bazie danych. "Inteligentne" identyfikatory mają jakieś

znaczenie. Na przykład ISBN jest dość inteligentnym identyfikatorem, gdyż poszczególne jego części niosą pewną informację. Jednakże w wypadku, gdy wydawca sprzeda prawa do książki innemu, inteligencja ISBN spada i dany ISBN może być tylko odnaleziony w centralnej bazie danych. Gdy mamy do czynienia z mniejszymi, często niesamodzielnymi jednostkami bibliograficznymi, wzrasta liczba potrzebnych identyfikatorów. Zarządzanie dużą ilością inteligentnych identyfikatorów stwarza poważne trudności, dlatego przeważa tendencja do używania raczej "głupich" identyfikatorów.

- Ważne jest rozróżnienie między identyfikatorem a lokalizacją. URL, który znany jest z powszechnego stosowania w WWW, jest bardziej lokalizatorem niż identyfikatorem. Gdy obiekt zmienia swoje miejsce, URL zmienia się również, a osoby używające poprzedniego URL mogą błędnie sądzić, że obiekt nie jest już dostępny (podczas gdy może on być dostępny w innej lokalizacji). Prawdziwy identyfikator musi pozostawać taki sam, bez względu na bieżącą lokalizację obiektu. Takie cechy ma, np. URN opracowany przez IETF (zob. niżej).
- Należy określić, jak bardzo niezmienny winien być identyfikator. W Internecie ogólnie przyjmuje się, że identyfikatory dla obiektów cyfrowych powinny trwać przez długi czas - znacznie dłuższy niż identyfikowane obiekty. Mają one więc przetrwać obecną technologię Internetu i systemów komputerowych.

Użycie identyfikatorów jako narzędzi daje wielkie możliwości. Przy ich pomocy można stwierdzić, że w obrębie jakiegoś obszaru intelektualnego dwa egzemplarze dzieła, którym przydzielono ten sam identyfikator są tym samym dziełem, podczas gdy dwa egzemplarze z różnymi identyfikatorami to dwa różne dokumenty. Jednak użycie identyfikatorów poza tymi ramami jest zawsze problematyczne. Na przykład zazwyczaj przydziela się różne ISBN w obrębie jednego wydania w zależności od rodzaju oprawy, dzięki czemu księgarnie mogą odróżnić wersje wydania, które różnią się także ceną i innymi warunkami nabycia. ISBN jest także wykorzystywany w opisach bibliograficznych; w tej sytuacji, gdy treść i paginacja wersji wydania w różnych oprawach jest taka sama lub równie dobrze spełni potrzeby czytelnika przeglądającego opisy, umieszczenie ISBN-u jako identyfikatora cytowanej pracy może spowodować problemy z powodu niepotrzebnego (dla tego zastosowania) rozróżnienia pomiędzy wersjami tego samego dzieła.

Identyfikatory uzyskały nowe znaczenie w środowisku sieci globalnej, gdzie procesy komputerowe pozwalają użytkownikowi na przejście od wystąpienia identyfikatora do uzyskania dostępu do identyfikowanego dzieła. Na przykład w Web można tworzyć linki pomiędzy pozycjami w bibliografii załącznikowej artykułu a cyfrowymi wersjami cytowanych prac. Linki te mogą być uruchamiane przy pomocy kliknięcia myszą. Znaczenie tak łatwego dostępu do cytowanych prac jest tak wielkie, że stało się przedmiotem kilku procesów sądowych. Jeżeli przez analogię przeniesie się te spory prawne na grunt dokumentów drukowanych, można zakwestionować prawa autora do cytowania innych prac bez zgody ich autorów. To, co jest normalną praktyką w publikacjach drukowanych, może zostać zabronione w środowisku sieciowym. Oczywiście jest to jedna z interpretacji tego zagadnienia, dodatkowo komplikowanego przez czynniki komercji. Pozwala ona na ukazanie, o jaką stawkę toczy się gra w tworzeniu systemów identyfikatorów, kontroli i praktyki ich używania.

W sieciowym środowisku zasobów cyfrowych obserwuje się powstawanie wielu ważnych identyfikatorów lub schematów przydzielania nazw[5], spośród których część już stosunkowo okrzepła, inne zaś są dopiero tworzone (np. URI, URN, DOI, Handles, ISBN, ISSN, SICI, BICI, ISWC, PII). Poniżej przedstawimy najważniejsze spośród tych inicjatyw. Jednak tylko niewiele z nich może funkcjonować w pełni efektywnie jako niezmienne identyfikatory ułatwiające dostęp do zasobów on-line w systemach rozproszonych. Do tego celu niezbędne jest zarejestrowanie ich jako schematu przydzielania nazw URI i wspomaganie przez system odwzorowania adresu (ang. resolution) lub włączenie do innego schematu nazw, który posiada jakąś formę związanego z nim systemu przekazywania.

System odwzorowania adresów w Internecie odgrywa podstawową rolę. Siła istniejącego standardu URL/URI wykorzystywanego w WWW polega na jego szeroko rozprzestrzenionym systemie odwzorowania adresów w formie DNS (Domain Name System). Bez systemu odwzorowania adresu zapytania używające identyfikatora nie mogą być przekazane do odpowiedniego serwera i wykorzystane do wyszukiwania źródła lub jego odpowiedniego substytutu w formie metadanych.

1. Uniwersalny Identyfikator Zasobów (Universal Resource Identifier - URI)

Web jest uniwersalną przestrzenią informacyjną w tym sensie, że obiekty w niej się znajdujące posiadają adresy. Te "adresy", "nazwy" lub jak się je powszechnie nazywa, identyfikatory, określono mianem Uniwersalnych Identyfikatorów Zasobów, w skrócie z angielskiego URI (*Universal Resource Identifier*).

Obiekt znajduje się w Web jeżeli posiada swój URI. Obiekty mające URI zwane są "Obiektami pierwszej klasy" (ang. *First Class Objects - FCO*). Web najlepiej działa, gdy mamy do czynienia z FCO. Jeżeli jakiś obiekt nie posiada URI, nie ma możliwości wykonania odesłania do niego, przez co maleje siła wyszukiwania w Web. Każde źródło znajdujące się w dowolnym miejscu może posiadać URI, a te najważniejsze powinny go posiadać. Oznacza to, że żadna informacja mająca jakiekolwiek znaczenie i trwałość nie możne być dostępna bez odniesienia do URI.

Wyróżnić można kilka zasad, którym podlega URI:

- **Globalny zakres**. Bez względu na to kto lub gdzie określił URI, będzie ono miało to samo znaczenie. Nie ma potrzeby określania miejsca, w którym URI musi być umieszczone aby mogło być utrzymywane. Wystarczy stwierdzenie, że obiekt jest "w Web". W powszechnym użyciu są linki rozpoczynające się od "http://", co oznacza użycie protokołu transferu hipertekstu. Jednak istnieje wiele innych URI, opisanych poniżej. Podstawą ich uniwersalizmu jest fakt, że identyfikatory te, bez względu na to, czy nazywane adresami czy lokalizatorami, są uniwersalne dzięki wykorzystaniu wielu wcześniej istniejących protokołów, takich jak SMTP i NNTP, protokołów specjalnie stworzonym dla Web (HTTP) oraz nowych, jeszcze nie stworzonych. Wynika z tego następna zasada: każda nowa nazwa lub adres może być przedstawiona jako podzbiór URI. Każdy nowy system nazewnictwa lub adresowania można łatwo włączyć przez kodowanie wartości nazw lub adresów na ciąg znaków i poprzedzenie standardowym prefiksem odpowiednim dla nowego schematu. Można więc zastąpić "http:" przez np. "ISBN:" lub "X500:", w zależności od wyboru najbardziej właściwego schematu identyfikatora.
- **Tożsamość**. URI ma w sposób powtarzalny odsyłać do tego samego obiektu. Ten sam ciąg znaków identyfikatora powinien bez względu na upływ czasu kierować do tego samego obiektu. Znaczenie identyfikacji dla danego URI określane jest przez osobę będącą jego właścicielem (tzn. tę, która pierwsza określiła, do czego on odsyła). Pojęcie właściciela zmienia się w zależności od przyjętego schematu. Jednak w każdym schemacie w praktyce określa się sposoby tworzenia nowych URI. W wielu schematach on sam powoduje lub wymaga wystąpienia pewnych własności identyfikacji i narzuca ograniczenia na właścicieli URI w definiowaniu identyfikacji. W efekcie będziemy potrzebowali protokołów dla otrzymywania potwierdzenia własności danego URI, nie są one bowiem zapisane w specyfikacji Web.
- **Brak wyłączności**. URI nie musi być jedynym uniwersalnym identyfikatorem. Może istnieć ich wiele, więc nie ma potrzeby walczyć o jeden mający specjalny status. Z drugiej strony zbyt wielka liczba identyfikatorów sprawia, że są one mało użyteczne.
- **Nieprzepuszczalność URI**. Identyfikator może być użyty jedynie do realizacji odesłania do obiektu. W większości wypadków użytkowania Web identyfikatory są wykorzystywane do tworzenia odesłań bez kontroli ich wewnętrznej zawartości. Nawet po uzyskaniu dostępu do identyfikowanego obiektu komputer klienta szuka pierwszej części identyfikatora wewnątrz HTTP, aby określić, z jaką maszyną serwera rozmawia. Następnie definiowana jest pozostała część ciągu niedostępnego dla klienta. Oznacza to, że klient nie sprawdza jego zawartości, nie może on wydedukować informacji ze znaków zawartych w identyfikatorze. Nieprzepuszczalność URI umożliwia tworzenie jego nowych schematów poprzez różne interpretacje HTTP URI. W efekcie wiele części metadanych, informacji o obiekcie, które mogłyby zostać wywnioskowane przez program klienta z bieżącej wartości ciągu URI, gdyby były w ten sposób dostępne, musi być udostępniana przez protokół HTTP.

Obecnie uważa się, że URI składa się z trzech różnych typów identyfikatorów: URL, URN i URC:

- **URL** są identyfikatorami określającymi fizyczną lokalizację, w której dostępne jest źródło, co oznacza, że zawiera on informację o tym, w jaki sposób źródło może być dostępne. Informacja ta zawiera określenie protokołu komunikacyjnego, nazwę komputera i nazwę pliku, w którym przechowywane jest źródło w formie cyfrowej.
- **URN** są symbolicznymi nazwami przydzielanymi źródłom przez autoryzowane organizacje. URN nie jest odnoszony do fizycznej lokalizacji, w której przechowywane jest źródło, a nazwa nie może zawierać żadnej informacji o samym źródle. Te wymagania są konsekwencją dążenia do używania URN jako długotrwałych odsyłaczy do źródła. URN ma realizować ideę długotrwałości i niezmienności (ang. *persistence*) w takim sensie, że URN raz nadany jest stale aktualny dla źródła, nawet w wypadku, gdy zmieniło ono lokalizację lub odpowiedzialność za źródło zostało przeniesione na inną organizację.
- **URC**, Uniform Resource Characteristics jest zestawem atrybutów zawierających dane o źródle, którego dotyczą. Traktowanie URC jako jednych z podstawowych URI pociąga za sobą potrzebę stosowania metadanych, które muszą określić i identyfikować, co jest najbardziej odpowiednim źródłem dla określonego użytkownika, w określonym kontekście, w danym czasie. Na przykład jeżeli źródło wskazane przez URN jest dostępne w kilku URL, URC może zostać użyty do automatycznego wybrania najbliższej lokalizacji. Może on też dostarczyć informacji o stopniu odpowiedniości źródła dla określonej grupy odbiorców (np. dzieci). Oczekuje się, że szersze wykorzystanie syntaktyki metadanych typu RDF spowoduje częstsze korzystanie z URC[6].

IETF używa terminu Uniform Resource Identifers (URI) jako nazwy wspólnej zarówno dla URL, jak i dla URN, wraz z dotąd nie zrealizowaną koncepcją URC, który może być rozumiany jako struktura pozwalająca powiązać jeden lub kilka URN z zestawami URL i metadanymi opisującymi obiekty identyfikowane przez te URN i URL. Przedstawione one zostały w RFC 2396 - Uniform Resource Identifiers (URI): Generic Syntax (Sierpień 1998).

2. OpenURL

Interesującym połączeniem mechanizmów wykorzystywanych w identyfikatorach internetowych i technikach metadanych jest projekt OpenURL [http://www.sfxit.com/openurl/]. Zawiera on mechanizmy służące do kodowania cytowania źródła informacyjnego, głównie opisu bibliograficznego, jako URL. W efekcie OpenURL jest

wykonywalnym URL, który przesyła metadane lub adresy umożliwiające do nich dostęp, dla obiektów, dla których utworzono OpenURL. OpenURL jest w zamierzeniu narzędziem służącym odesłaniom (ang. *Resolver*[7]) do źródeł poprzez prowadzenie wyszukiwania rzeczowego w oparciu o metadane.

Opis bibliograficzny jest udostępniany albo przy pomocy globalnego identyfikatora dla źródła takiego jak DOI, albo przez pobranie istniejących metadanych o źródle lub przez połączenie obu możliwości. Możliwe jest także kodowanie lokalnego identyfikatora źródła w OpenURL. Posiadanie informacji o miejscu utworzenia OpenURL pozwala oprogramowaniu, które natrafiło na OpenURL na ściągniecie metadanych o źródle informacji.

Open URL składa się z dwóch części, BASEURL i QUERY. BASEURL identyfikuje resolver OpenURL, który będzie realizował usługi dotyczące treści dokumentów dla Open URL. BASEURL jest określony dla konkretnego użytkownika, który wysyła OpenURL - identyfikuje on resolver OpenURL, który jest preferowany przez użytkownika. W wielu wypadkach będzie to resolver oferowany przez instytucję, do której przynależy użytkownik. Usługi realizowane przy pomocy OpenURL dotyczące interfejsów Web, np. sposobu wyświetlania wyników wyszukiwania, muszą stworzyć mechanizmy dla powiązania BASEURL z każdym użytkownikiem końcowym. Jedną z możliwości osiągnięcia tego celu jest zapisanie BASEURL w cookie[8] w wyszukiwarce użytkownika, innym jest zapisanie BASEURL wraz z innymi preferencjami użytkownika.

Część QUERY może składać się z jednej lub kilku DESCRIPTION. Każdy DESCRIPTOIN zawiera atrybuty i wartości metadanych, które tworzą opis źródła.

W artykule Encoding OpenURLs in Dublin Core metadata, Andy Powell i Ann Apps przedstawili następujące przykłady OpenURL:

> **http://resolver.ukoln.ac.uk/openresolver/?sid=ukoln:ariadne&genre=article**
> **&atitle=Information%20gateways:%20collaboration%20on%20content**
> **&title=Online%20Information%20Review&issn=1468-4527&volume=24**
> **&spage=40&epage=45&artnum=1&aulast=Heery&aufirst=Rachel**[9]

W podanym przykładzie częścią BASEURL jest **http://resolver.ukoln.ac.uk/openresolver/** - jest to URL wersji demonstracyjnej OpenResolver przygotowanego w UKOLN. Pozostała część OpenURL to QUERY, składająca się z jednego DESCRIPTION artykułu zatytułowanego "Information gateways: collaboration on content" autorstwa Rachel Heery. Artykuł ten opublikowany został w Online Information Review nr 24.

Wybór przez użytkownika OpenURL jakiegoś dokumentu przez kliknięcie oznacza żądanie dostarczenia rozszerzonych usług dla tego dokumentu. Jako dane wejściowe pobierany jest OpenURL oraz gromadzone są metadane i identyfikatory dla tego dokumentu. Informacja ta może być bezpośrednio odczytana z OpenURL lub przeniesiona przy zastosowaniu wskaźnika metadanych użytego w OpenURL. Wskaźnik ten może prowadzić do pierwotnego lub do jakiegokolwiek innego źródła. Po zgromadzeniu metadanych i identyfikatorów zostaną one ocenione. Następnie użytkownikowi przedstawione zostaną rozbudowane linki. Jeżeli usługa jest odpowiednio przystosowana do indywidualnych potrzeb, linki te będą odpowiadały wymogom użytkownika. Ilość i jakość dostępnych metadanych odgrywa zasadniczą rolę w sposobie realizacji i jakości usług dostępnych dla danego obiektu.

OpenURL począwszy od 2000 roku jest stosowany coraz częściej. Został on zaimplementowany przez takich dostawców zasobów informacyjnych jak ISI, Ebsco, Silver Platter, czy Swets. Jednym z pierwszych zintegrowanych systemów bibliotecznych, które w pełni zastosowały ten standard jest Aleph500 firmy Ex Libris, działający także w polskich bibliotekach. Istotne znaczenie mają także prace nad połączeniem działań związanych z tworzeniem OpenURL oraz DOI. Po zapowiadanym zatwierdzeniu standardu NISO dla OpenURL liczba dalszych implementacji z pewnością wzrośnie.

3. Persistent URL (PURL) z OCLC

W odpowiedzi na problem niestałości adresowania przy pomocy URL, OCLC przygotowało system nazwany PURL (Persistent URL). Zasadniczo PURL jest to URL wyrażony w HTTP, gdzie nazwa hosta zastąpiona została przez host "PURL.ORG" a nazwa pliku jest identyfikatorem dla "rzeczywistych" treści, do których odsyła identyfikator. Host PURL.ORG będzie utrzymywany długoterminowo pod tą nazwą; jeżeli ktoś zarejestruje obiekt na serwerze PURL, otrzymuje on bieżącą nazwę hosta i pliku dla obiektu, a serwer PURL tworzy pozycję bazy danych łączącą nazwy hosta i pliku z identyfikatorem, który będzie znajdował się w PURL. Gdy dany PURL jest wywoływany (zostaje użyty), następuje połączenie z serwerem PURL, wyszukanie identyfikatora w jego bazie danych, sprawdzanie, gdzie poszukiwany obiekt obecnie się znajduje. Następnie wykorzystuje się możliwości przekierowywania protokołu HTTP, aby połączyć poszukującego z hostem, na którym znajduje się obiekt. Administratorzy obiektów sieciowych są odpowiedzialni za przesyłanie danych o modyfikacjach lokalizacji do serwera PURL, gdzie zmieniane są nazwa pliku i/lub jego lokalizacja.

PURL reprezentuje ideę pośredniego ustalania adresów - wyszukuje identyfikator w bazie danych, aby sprawdzić,

gdzie obiekt aktualnie się znajduje. Jest to bardzo pomysłowy i praktyczny sposób, gdyż współpracuje się z istniejącą, zainstalowaną bazą wyszukiwarek Web. Jednak nie używa się prawdziwych nazw, gdyż pozwalają one na dostęp do obiektu poprzez określoną usługę, tzn. HTTP. PURL nie będzie prawdopodobnie dalej funkcjonował, gdyż pojawiają się nowe protokoły, zastępujące HTTP.

4. Handle System i DOI

System Handle® (http://www.handle.net/) jest systemem nazewnictwa ogólnego stosowania, który pozwala na podejmowanie bezpiecznych decyzji i administrowanie w zakresie nazewnictwa w obrębie ogólnie dostępnego Internetu[10]. Projekt ten udostępnia zestaw protokołów, przestrzeni nazw oraz implementacji tych protokołów. Protokoły umożliwiają odległym systemom komputerowym na zachowanie nazw źródeł cyfrowych i przekształcenie tych nazw w informację niezbędną dla zlokalizowania, udostępniania i innego rodzaju korzystania ze źródła. Możliwa jest taka zmiana związanych danych, aby odzwierciedlały one aktualny stan identyfikowanego źródła bez konieczności zmiany nazwy. Pozwala to na niezmienność nazwy pozycji pomimo zmian lokalizacji i innych informacji o stanie bieżącym. Każda nazwa może mieć własnego administratora; administrowane one mogą być też w sposób zdalny poprzez sieć. Możliwe jest także zastosowanie procedur zabezpieczających, co pozwala na użycie systemu w zastosowaniach, w których wiarygodność odgrywa dużą rolę.

Handle System wymyślono i rozwinięto w CNRI (Corporation for National Research Initiatives, USA) jako część projektu Computer Science Technical Reports (CSTR) finansowanego przez DARPA (Defense Advanced Projects Agency). Jednym z ważniejszych osiągnięć tego projektu związanego z bibliotekami cyfrowymi, który wpływał na ewolucję takich prac jak NCSTRL (Networked Computer Science Technical Reference Library - http://www.ncstrl.org/), było stworzenie podstaw dla ogólnej struktury bibliotek cyfrowych. Pierwsza implementacja zrealizowana w CNRI została udostępniona w Internecie w końcu roku 1994.

Jednymi z pierwszych użytkowników Handle System byli Library of Congress i Indernational DOI Foundation. Współpraca z tymi instytucjami wpłynęła na dalszy rozwój systemu.

Handle System zaprojektowany został w taki sposób, aby można było pominąć ograniczenia najpowszechniej używanych systemów nazewnictwa w Internecie: DNS i URL, a także zwiększyć funkcjonalność nazw. Handle System zapewnia realizację następujących celów:

- **Jednoznaczność**. Każda nazwa jest unikalna w skali globalnej w obrębie Handle System.
- **Długotrwałość**. Nazwa nie jest pobierana w żaden sposób z pozycji którą nazywa, ale jest przydzielana jej w sposób niezależny. Dzięki temu, że do Handle można włączyć istniejącą (nawet mnemoniczną) nazwę, przez Handle System obsługiwane jest jedyne działające połączenie pomiędzy nazwą i pozycją, która jest nazywana. Nie gwarantuje to oczywiście długotrwałości, która jest raczej efektem staranności w administrowaniu nazwami, ale pozwala na wykorzystywanie tej samej nazwy pomimo, np. zmian lokalizacji czy własności. Jeśli nazwane źródło przenoszone jest z jednej lokalizacji do innej, nazwa może pozostać aktualna dzięki aktualizacji jej wartości odzwierciedlającej w Handle System zmianę lokalizacji.
- **Wiele wystąpień**. Pojedyncza nazwa może odnosić się do wielu wystąpień źródła w różnych i zmiennych lokalizacjach w sieci. Na przykład usługi sieciowe mogą określać wiele punktów dostępu do tych usług z wykorzystaniem jednej nazwy, w związku z czym następuje rozproszenie dostępu i zmniejszenie obciążenia.
- **Rozszerzalność przestrzeni nazw**. Istniejące lokalne przestrzenie nazw mogą być dołączane do przestrzeni nazw Handle w wyniku otrzymania unikalnej nazwy poprzez autoryzację nazw. Pozwala to lokalnej przestrzeni nazw uzyskać zasięg światowy przy jednoczesnym uniknięciu konfliktu z istniejącymi przestrzeniami nazw. Autoryzacja nazw pozwala także na przekazanie realizacji usług zarówno decyzyjnych, jak administracyjnych do lokalnych usług Handle.
- **Zasięg międzynarodowy**. Przestrzeń nazw Handle jest oparta na Unicode 2.0, który zawiera większość znaków diakrytycznych używanych obecnie w świecie, co pozwala na wykorzystanie systemu w każdym kraju. Protokół Handle używa UTF-8[11] dla kodowania nazw.
- **Rozproszony model usług**. Handle System definiuje hierarchiczny model usług w taki sposób, że każda lokalna przestrzeń nazw może być obsługiwana albo przez odpowiednie lokalne usługi nazewnictwa lub przez usługi globalne, bądź przez obie. Usługi globalne, znane pod nazwą Global Handle Registry mogą być wykorzystywane do wysyłania każdego zapytania do usług Handle do odpowiedniego lokalnego serwisu Handle. Rozproszony model usług pozwala na powielenie dowolnych usług w wielu miejscach.
- **Bezpieczne usługi nazewnictwa**. Protokół Handle pozwala na autoryzację klientów przez serwery i realizację usług integracji danych na życzenie klienta. Możliwe jest użycie kodowanych kluczy dostępu. Ta cecha może być wykorzystywana do uniemożliwienia podsłuchiwania i fałszowania przysyłanych zapytań klienta lub przechwytywania odpowiedzi serwera.
- **Rozproszone usługi administracyjne**: Każda nazwa może posiadać własnego administratora lub grupę administrującą. W połączeniu z protokołem autoryzacji pozwala to na bezpieczną obsługę nazw poprzez sieć publiczną przez autoryzowanych administratorów w dowolnej lokalizacji w sieci.

Każda nazwa składa się z dwóch części: nazwy instytucji autoryzowanej (ta część nazwy określana jest mianem

prefiksu) oraz unikalnej nazwy lokalnej z zakresu autoryzowanego nazewnictwa, zwana sufiksem. Instytucja autoryzowana i nazwa lokalna są oddzielone znakiem ASCII "/". Tak więc nazwa może być przedstawiona jako schemat:

<Handle>::=<Instytucja autoryzowana>"/"<Nazwa lokalna>

Tak więc np. "10.1045/january99-bearman" jest nazwą dla artykułu opublikowanego w czasopiśmie D-Lib. Czasopismo jest zdefiniowane pod nazwą instytucji autoryzowanej "10.1045" a nazwa lokalna dla artykułu brzmi "january99-bearman".

System Handle wykorzystuje obiekty zwane handles w celu określenia lokalizacji i innych informacji o źródle. Składa się on z następujących części:

- Instytucje nadające nazwy. Są to jednostki administracyjne posiadające autoryzację do tworzenia i obsługi handles. Nazwa instytucji nadającej nazwy zawiera jeden lub więcej ciągów, oddzielonych kropką.
- Serwery Handle. Tu handles są zapisywane i realizowane są usługi ich przenoszenia. Przejścia dla handles są realizowane przez serwery handle na żądanie klienta. Istnieje jeden Global Handle Server lub Registry i wiele podległych instytucji nadających nazwy.
- Serwery cache'ujące[12]. Są związane z lokalnymi serwerami, pozwalają na przenoszenie często udostępnianych handles bez potrzeby pobierania adresu z Global Registry.
- Serwery proxy pozwalają użytkownikom Web i innym klientom na przenoszenie handles.

Przestrzeń nazw handle może być uważana za zestaw wielu lokalnych przestrzeni nazw, w której każda lokalna przestrzeń nazw posiada własną unikalną instytucję przydzielającą nazwy. Instytucja ta identyfikuje jednostkę administrującą tworzeniem nazw powiązanych. Każda instytucja nazywająca jest unikalna w skali globalnej w obrębie Handle System.

W 1997 roku Association of American Publishers (AAP) oraz współpracująca z nim Corporation for National Research Initiatives (CNRI) ogłosiły powstanie nowego identyfikatora nazwanego Digital Object Identifier (DOI). DOI oparty jest na The Handle System utworzonym przez CNRI. Jest to bardzo ogólny system identyfikacji, który zawiera się z grubsza w ramach URN i który dostarcza mechanizmów dla implementacji systemu nazw dla obiektów cyfrowych. Obecnie sprawami DOI zarządza International DOI Foundation - IDF (http://www.doi.org/). DOI został utworzony dla umożliwienia zarządzania prawami własności (copyright) w środowisku elektronicznym. Jego autorzy wyszli z założenia, że aby coś chronić, najpierw należy w sposób niepowtarzalny i jednoznaczny określić czym jest chroniona jednostka. Dlatego też prace nad DOI rozpoczęto jako inicjatywę mającą praktyczne cele, mianowicie przygotowanie zasad jednoznacznego, niezmiennego nazewnictwa. Jest to pierwsza część większej implementacji zawierającej narzędzia pozwalające na zarządzanie jednoznacznie nazwanymi jednostkami. Ujmując rzecz najprościej system DOI oferuje:

- niezmienny identyfikator własności intelektualnej; oraz
- mechanizm dla przetworzenia tego identyfikatora na pewne przydatne informacje lub usługi.

Drugi z tych dwóch punktów wiąże identyfikator z pewną określoną informacją; w najprostszy sposób powiązanie to jest realizowane przez odesłanie do strony Web wydawcy. Wówczas DOI staje się wskaźnikiem przejścia do wystąpienia materiału na stronie Web wydawcy. Wiele prototypowych zastosowań DOI realizuje tę funkcję. Jednak nie zamierza się na tym poprzestać, gdyż ograniczenie użytkowania identyfikatora tylko do łączenia się ze stronami wydawców nie jest ani właściwe, ani praktyczne. Intencją twórców DOI jest aby był on publicznym identyfikatorem używanym w wielu zastosowaniach, także do użytku lokalnego, oraz, jak inne identyfikatory informacji, DOI powinien być niezależny od konkretnych zastosowań - powinien to być identyfikator, którego wielu użytkowników może swobodnie używać w wielu aplikacjach[13].

DOI dostarcza niezmiennych identyfikatorów ("nazw") dla źródła lub pozycji. Pozwala to na bezpośredni dostęp do pozycji, w odróżnieniu od URL, który określa lokalizację, w której znajduje się kopia pozycji. Stąd DOI pozwala na stworzenie infrastruktury umożliwiającej zarządzanie obiektami cyfrowymi niezależnie od ich lokalizacji.

Ogólną formą URN jest **urn:nid:nss**, gdzie **nid** oznacza zdefiniowany identyfikator (np. doi), a **nss** jest ciągiem określającym nazwę w obrębie tego nid. Koncepcja URN pozwala na opcjonalne włączenie identyfikatora schematu nazwy (si), jako **urn:si:nid:nss**. W tym przypadku czytanie nazwy od lewej do prawej strony jest równoznaczne z przechodzeniem w dół w hierarchicznej strukturze nazwy. Na przykład możemy przedstawić typowy DOI:

urn:doi:10.1000/123456789

W tym przykładzie "doi" jest identyfikatorem typu nazwy (nid); 10.1000/123456789 jest ciągiem określającym nazwę (nss). Jeżeli dane o typie identyfikatora wynikają z kontekstu, jest łatwiej zapisywać tylko ciąg określający nazwę i odnosić się do niego jako DOI. W powyższym przykładzie DOI jest dzielony na dwie części przez "/". Część przed "/" (10.1000) jest nazywana prefiksem i jest administrowana przez organizację tworzącą i zarządzającą DOI; ta organizacja jest z kolei tworzona przez agencję posiadającą uprawnienia do DOI (obecnie International DOI Foundation).

Syntaktyka DOI jest syntaktyką Handle:

<Kod katalogu>.<Kod rejestru>/<Sufiks DOI>

DOI jest więc ciągiem znaków alfanumerycznych, wśród których wyróżnia się prefiks i sufiks. Prefiks identyfikuje pierwotną agencję rejestrującą i rejestratora, natomiast sufiks zawiera indywidualny identyfikator produktu, poprzedzony kodem oznaczającym rodzaj użytego identyfikatora. Kod katalogu dotychczas zawsze dla DOI przyjmuje wartość "10". Kod rejestru przydzielany jest przez IDF określonym wydawcom, właścicielom praw autorskich i instytucjom rejestrującym. Sufiks DOI może być dowolnym ciągiem odpowiadającym wymaganiom instytucji rejestrującej.

Innym przykładem DOI może więc być:

10.1234/[ISBN] 83 87629 84 7.

Jak zatem widać szczegółowy sposób użycia DOI zależny jest od typu dokumentu, dla jakiego identyfikator został wykorzystany. Należy podkreślić, że część symbolu przed kreską ukośną oznacza pierwotną agencję i pierwotnego rejestratora. Jeżeli rejestrujący sprzeda rejestrowany produkt, DOI produktu nie ulegnie zmianie, nawet jeżeli nowy rejestrujący jest znany. Dlatego też pierwotne znaczenie symbolu DOI jest tracone po zmianie właściciela. Jednak z technicznego punktu widzenia katalog, w którym DOI jest rejestrowany, będzie automatycznie poprawnie przekierowywać zapytania do nowego właściciela.

Można stwierdzić, że DOI spełnia wymagania stawiane URN i ma możliwości sprostania wszystkim wymogom opisanym wcześniej. DOI nie tylko umożliwia tworzenie niezmiennych nazw (identyfikatorów), ale także wykorzystuje system odwzorowania adresów używając technologii wspomnianego systemu - Handle System (http://www.handle.net/). Handle System[14] można uznać za zgodny z URN. Oferuje on dodatkowe możliwości i cechy, w tym kilka jeszcze nie zaimplementowanych w DOI.

Rekord DOI wraz z bieżącą lokalizacją obiektu identyfikowanego jest rejestrowany na serwerze systemu DOI. Serwer ten funkcjonuje jako resolwer do obiektów, które znajdują się na stronach wydawców lub stronach przez nich licencjonowanych. Jak już stwierdziliśmy, obecnie, dla pierwszych implementacji większość DOI odsyła do strony wydawcy, jednak pełna realizacja systemu DOI przez utworzenie możliwości wielokrotnego przejścia i zastosowanie metadanych zmieni tą sytuację. Zmiany lokalizacji są rejestrowane na centralnym serwerze.

DOI jest coraz szerzej wykorzystywany w środowisku informacji elektronicznej, głównie przez wydawców dokumentów elektronicznych. Przykładem może być inicjatywa o nazwie CrossRef (http://www.crossref.org/), w ramach której współpracuje większość wydawców publikacji naukowych. CrossRef działa jak rodzaj przełącznika cyfrowego. Zgromadzone są w nim opisy dokumentów, które pozwalają wydawcom i innym współpracującym organizacjom na tworzenie linków do pełnych tekstów dokumentów przy użyciu DOI działającego jak etykieta artykułu w metadanych dostarczanych przez wydawcę. Wybór linku poprzez kliknięcie spowoduje połączenie ze stroną Web wydawcy na której znajduje się pełny opis bibliograficzny poszukiwanego artykułu, oraz w większości przypadków jego abstrakt. Następnie czytelnik może uzyskać dostęp do pełnego tekstu artykułu poprzez specjalny mechanizm, który selekcjonuje użytkowników na tych, którzy prenumerują czasopismo (pełny, bezpłatny dostęp) i innych, którzy mogą uzyskać dostęp płatny.

Pomimo, że DOI wciąż nie jest głównym identyfikatorem dla zasobów Web, jego syntaktyka została zaaprobowana przez amerykańską agencję NISO w maju 2000 roku i opublikowana jako oficjalny standard - Z39.840-2000.

Przypisy

[1] COUSINS, Shirley. Duplicate detection and record consolidation in large bibliographic databases : the COPAC database experience. *Journal of Information Science*, 1998, vol. 24, no. 4, pp. 231-240.

[2] Dack, Diana. Persistent identification systems In *National Library of Australia* [on-line]. [dostęp 21 października 2003]. Dostępny w World Wide Web: http://www.nla.gov.au/initiatives/persistence/PIcontents.html.

[3] HAKALA, Juha. Principles of identification : European perspectives. In International Conference Electronic Resources : Definition, Selection and Cataloguing : 26th 27th 28th November 2001 [online]. [dostęp 21 października 2003]. Dostępny w World Wuide Web: http://w3.uniroma1.it/ssab/er/relazioni/hakala_eng.pdf.

[4] NAHOTKO, Marek. Nowe potrzeby w zakresie identyfikacji. *Bibliotekarz*, 2003, nr 4, s. 13-16.

[5] Dack, Diana [Dok. elektr.] (2001). Persistent identification systems. Dack, Diana. Persistent identification systems In *National Library of Australia* [on-line]. [dostęp 21 października 2003]. Dostępny w World Wide Web: http://www.nla.gov.au/initiatives/persistence/PIcontents.html.

[6] CUENCA, Pedro, SOSA, Vincente. Experiences in the use of metadata for web publishing. In ICCC/IFIP Third

Conference on Electronic Publishing '99 [on-line]. [dostęp 21 października 2003]. Dostępny w World Wide Web: http://www5.hk-r.se/elpub99/ap.nsf/08c6c2f88424ad99c12566ff002a0c10/c9af95de1c8ab4c9c12567980042f32b!OpenDocument

[7] Resolver - program-klient komunikujący się z serwerami nazw. Jeśli znamy nazwę komputera, z którym chcemy się połączyć, połączenie można zrealizować nawet po zmianie adresu IP tego komputera, ponieważ nazwy domen są tłumaczone na adresy IP "na bieżąco". Program określający translację adresów z postaci liczbowej do symbolicznej.

[8] Cookie - tak nazywa się małą (cookie oznacza ciasteczko) porcję (pakiet) informacji przesyłaną przez serwer do przeglądarki (ogólniej - do klienta) i przechowywaną w pamięci, a także zapisywaną na dysku klienta. Przy kolejnym połączeniu z tym samym serwerem może to przyspieszyć obsługę, np. dzięki znajomości preferencji klienta.

[9] POWELL, Andy, APPS, Ann. Encoding OpenURLs in Dublin Core metadata. In *Ariadne* [on-line]. 2001 iss 27.]. [dostęp 21 października 2003]. Dostępny w World Wide Web: http://www.ariadne.ac.uk/issue27/metadata/. ISSN 1361-3200.

[10] LANNOM, Laurence. Handle System Overview. In *ICSTI : Welcome to ICSTI* [on-line]. [dostęp 21 października 2003]. Dostępny w World Wide Web: http://www.icsti.org/icsti/forum/fo9904.html.

[11] UTF-8, A Transformation Format for Unicode and ISO 10646. In *IETF The Internet Engineering Task Force* [on-line]. [dostęp 21 października 2003]. Dostępny w World Wide Web: http://www.ietf.org/rfc/rfc2044.txt.

[12] Nazwa cache oznacza tu miejsce na dysku twardym, w którym przeglądarka internetowa przechowuje odwiedzone uprzednio strony (lub części stron) WWW. Ma to na celu przyspieszenie przeglądania stron internetowych, gdyż nie ma potrzeby ściągania odwiedzonych już stron z Internetu.

[13] Na przykład: The Copyright Clearance Center (CCC) w USA przydzielił DOI dla cyfrowych obrazów fotografii ze swoich zbiorów liczących 50.000 egz. Klikniecie na DOI spowoduje przesłanie zapytania do CCC, która reprezentuje fotografików będących właścicielami praw autorskich tych cyfrowych dokumentów.

[14] Handle System : *Handle Resolution* [on-line]. [dostęp 21 paździrnika 2003]. Dostępny w World Wide Web: http://www.handle.net/overviews/hs-version4.html.

(C) 2003 EBIB

Identyfikacja obiektów w sieciach rozległych / Marek Nahotko // W:Internet w bibliotekach II [Dokument elektroniczny] : łączność, współpraca, digitalizacja : Wrocław, 23-26 września 2003 roku. - Dane tekstowe. - [Warszawa] : Stowarzyszenie Bibliotekarzy Polskich, K[omisja] W[ydawnictw] E[lektronicznych], Redakcja "Elektronicznej Biblioteki", 2003. - (EBIB Materiały konferencyjne). - Tryb dostępu : http://ebib.oss.wroc.pl/matkonf/iwb2/nahotko.php . - Internet w bibliotekach II. - ISBN 83-915689-5-4