5. Trwała identyfikacja publikacji w repozytoriach cyfrowych – przegląd stosowanych systemów

Aneta Januszko-Szakiel

Wprowadzenie

Publikowanie w internecie oraz tworzenie repozytoriów, w których gromadzi się i archiwizuje cyfrowe kolekcje różnorodnych treści, stało się zjawiskiem powszechnym. Istotne atrybuty repozytoriów cyfrowych to przede wszystkim długoterminowa, niekiedy wieczysta archiwizacja oraz jednoznaczna identyfikacja i wyszukiwanie przechowywanych w nich obiektów[1]. Szczególnego znaczenia atrybuty te nabierają w przypadku repozytoriów bibliotecznych, archiwalnych, uczelnianych *etc.*, których obiekty stanowią narodowe dziedzictwo cyfrowe i służą jako zaplecze wiedzy w procesach edukacyjnych, pracach naukowych i badawczych. Ich dostępność oraz czytelność powinna być zagwarantowana pomimo wszelkich technologicznych i organizacyjnych zmian, m.in. poprzez jednoznaczne adresowanie i identyfikowanie. Jeżeli nie zostanie zagwarantowana zarówno dostępność, jak i czytelność publikacji sieciowych, wówczas użyteczność zasobów repozytoriów będzie znacznie ograniczona, np. poprzez brak możliwości cytowania i odwoływania się do treści tych dokumentów.

W przypadku publikacji tradycyjnych system identyfikowania jest powszechnie znany. Polega na przydzielaniu publikacjom znormalizowanych, jednoznacznych i niepowtarzalnych numerów ISBN, ISSN, ISAN bądź ISMN[2].

---

[1] W niniejszym opracowaniu przez pojęcie obiektu w repozytorium cyfrowym należy rozumieć pojedynczy dokument opublikowany w sieci. W zależności od typu repozytorium obiektem może być opublikowana w formie elektronicznej książka, artykuł, rozprawa doktorska, habilitacyjna, baza danych, prezentacja PowerPoint, nagranie wykładu, także inne formy prezentacji treści zapisane w postaci kodu zerojedynkowego. Wraz z terminem *obiekt* zamiennie występują pojęcia: materiał cyfrowy, zasób cyfrowy, dokument cyfrowy, publikacja cyfrowa, obiekt sieciowy.

[2] ISBN – International Standard Book Number (Międzynarodowy Znormalizowany Numer Książki), ISSN – International Standard Serial Number (Międzynarodowy Znormalizowany Numer Wydawnictwa Ciągłego), ISAN – International Standard Audiovisual Number (Międzynarodowy Znormalizowany Numer Utworów Audiowizualnych), ISMN – International Standard Music Number (Międzynarodowy Znormalizowany Numer Druku Muzycznego). Szczegółowe informacje o identyfikatorach dokumentów zamieszcza w swoim serwisie WWW Biblioteka Narodowa: http://www.bn.org.pl/index.php [dostęp: 12.01.2009].

Podobne systemy identyfikacyjne są stosowane dla obiektów sieciowych. W procesach bibliograficznych odesłań i wyszukiwania obiektów sieciowych posługiwanie się tylko obiegowymi adresami internetowymi URL (Uniform Resource Locators) jest niewystarczające, gdyż te zmieniają się zbyt często. Profesjonalne repozytoria cyfrowe, dbające o użyteczność[3] zdeponowanego materiału cyfrowego stosują rozmaite systemy trwałego identyfikowania obiektów.

W niniejszym rozdziale zdefiniowano pojęcie „repozytorium cyfrowe" oraz dokonano przeglądu powszechnie stosowanych systemów trwałej identyfikacji obiektów sieciowych.

5.1. Definicja repozytorium cyfrowego

Z przeglądu definicji dostępnych w piśmiennictwie przedmiotu[4] wynika, że przez pojęcie „repozytorium elektroniczne" tudzież „repozytorium cyfrowe" należy rozumieć organizację ludzi oraz narzędzi lub system złożony z osób oraz przyjętych rozwiązań organizacyjnych i technicznych, powołany w celu zgromadzenia, przechowania oraz zapewnienia długoterminowego dostępu i użyteczności cyfrowego materiału. Działania repozytorium koncentrują się na pracach związanych z przeprowadzeniem cyfrowych dokumentów przez kolejne etapy rozwoju technologicznego, przy użyciu najróżniejszych narzędzi i metod archiwizacji, między innymi migracji oraz emulacji[5]. Docelowo repozytorium ma dostarczyć obecnym oraz przyszłym

[3] Przez pojęcie użyteczności cyfrowych zasobów archiwalnych należy rozumieć m.in. stabilny dostęp do autentycznych i integralnych dokumentów cyfrowych oraz możliwość powoływania się na nie we własnych opracowaniach poprzez stosowanie bibliograficznych odesłań. Źródło: *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources. RLG-OCLC Report*, Mountain View, CA, August 2001, [online:] http://www.rlg.org/longterm/attributes01.pdf [dostęp: 20.12.2008], A. Januszko-Szakiel, *Archiwizacja publikacji elektronicznych jako wyzwanie dla bibliotek – zarys problematyki*, „Biuletyn Biblioteki Jagiellońskiej" 2003, s. 216–225.

[4] *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources. RLG-OCLC Report*. Mountain View, CA. August 2001, [online:] http://www.rlg.org/longterm/attributes01.pdf [dostęp 20.12.2008]; G. Clavel-Merrin, *The Nedlib List of Terms. Nedlib Report Series 7*, Amsterdam 2000, s. 3; *Kriterienkatalog vertrauenswürdige digitale Langzeitarchive. Version 1. (Entwurf zur öffentlichen Kommentierung). Nestor Materialien 8*. Frankfurt am Main, 2006, [online:] http://edoc.hu-berlin.de/series/nestor-materialien/2006-8/PDF/8.pdf, s. 2 [dostęp: 20.08.2008]; J.M. Reitz, *Dictionary for Library and Information Science*, Westport–London 2004, s. 216.

[5] U.M. Borghoff i in., *Langzeitarchivierung. Methoden zur Erhaltung digitaler Dokumente*, Heidelberg 2003, s. 47–78; A. Januszko-Szakiel, *Rola migracji i emulacji w strategii długoter-*

użytkownikom możliwość odczytu autentycznych, integralnych, wiarygodnych i poufnych dokumentów cyfrowych[6].

W wypowiedziach na temat repozytoriów cyfrowych autorzy często odwołują się do standardu archiwizacji publikacji elektronicznych OAIS, w którym oprócz wymienionych cech uwzględnia się dążenie repozytorium cyfrowego do stałej obserwacji i zabezpieczenia zmieniających się potrzeb docelowej grupy użytkowników, nazywanych niekiedy klientami, tudzież odbiorcami usług repozytorium. Synonimicznie „repozytorium cyfrowe" określane bywa terminem „archiwum cyfrowe"[7]. W dalszej części tekstu terminy „repozytorium" oraz „archiwum cyfrowe" bądź „archiwum elektroniczne" będą stosowane wymiennie.

Model referencyjny repozytoriów cyfrowych OAIS został stworzony przez *Consultative Committee for Space Data Systems* (CCSDS)[8] na potrzeby archiwizacji i wymiany danych elektronicznych, zawierających informacje z badań przestrzeni kosmicznej. W maju 1999 r. zaprezentowana została pierwsza wersja modelu OAIS, a w lutym 2003 r., po licznych poprawkach model OAIS został zaakceptowany przez International Organization for Standarization jako norma postępowania w zakresie długoterminowej archiwizacji danych cyfrowych (ISO 14721:2003).

Pomimo że model OAIS został stworzony głównie z myślą o archiwizacji jednego typu danych elektronicznych, jest on uznawany za uniwersalny model organizowania i funkcjonowania repozytoriów cyfrowych i stosowany do gromadzenia, przechowywania i udostępniania różnych typów dokumentów elektronicznych. OAIS jest wykorzystywany w wielu światowych bibliotekach, archiwach i muzeach, w których realizowane są projekty długoterminowej archiwizacji zbiorów cyfrowych.

Jednym z kluczowych pojęć w modelu referencyjnym OAIS jest pakiet informacyjny – *Information Package*. Składa się on z dwóch komponentów, tj.

---

minowej archiwizacji publikacji elektronicznych, [w:] *Informatyka*, red. M. Pękala, W.Z. Chmielowski, Kraków 2008, s. 121–130.

[6] *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Nestor Handbuch*, 2008, [online:] http://nestor.sub.unigoettingen.de/handbuch/nestor-handbuch.pdf [dostęp: 20.08.2008]; A. Januszko-Szakiel, *Archiwizacja publikacji elektronicznych jako wyzwanie dla bibliotek – zarys problematyki.* „Biuletyn Biblioteki Jagiellońskiej" 2003, s. 215–220.

[7] J.M. Reitz, *Dictionary for Library and Information Science*, Westport–London 2004, s. 216; *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*, 2002, [online:] http://www.rlg.org/longterm/repositories.pdf [dostęp: 20.12.2008].

[8] Komitet CCSDS został powołany w 1982 r. Jest organizacją składającą się z przedstawicieli wielu światowych agencji badań przestrzeni kosmicznej i podlega bezpośrednio agencji NASA; szczegółowe informacje zob. http://www.ccsds.org/ [dostęp: 06.01.2009].

kontenera informacyjnego (*Content Information*) oraz informacji dotyczącej przechowywania jego zawartości (*Preservation Description Information* – PDI). PDI to w myśl modelu OAIS wszelkie informacje konieczne do odpowiedniego przechowania informacji treściowej (kontenera informacyjnego). Zalicza się tu cztery typy informacji, określane jako: historia (*Provenance*), powiązania (*Context*), identyfikatory (*Reference*) oraz mechanizmy ochrony danych – *Fixity*.

- *Provenance*, w dosłownym tłumaczeniu „pochodzenie", określa źródło obiektu informacyjnego, wskazuje na podmiot odpowiedzialny za opiekę nad obiektem od momentu jego powstania oraz dostarcza wiedzy na temat historii obiektu.
- *Context* opisuje związek obiektu informacyjnego z innymi obiektami nienależącymi do danego pakietu informacyjnego.
- *Reference* jest odpowiedzialny za dostarczenie identyfikatorów, umożliwiających jednoznaczną identyfikację obiektu informacyjnego. Najogólniej rzecz ujmując, zadaniem identyfikatorów publikacji elektronicznych jest odróżnienie określonej publikacji od innych. W archiwach elektronicznych identyfikatory występują pod nazwą *Digital Object Identifier* (DOI) czy też *Persistent Identifier* (PI).
- *Fixity* to element wprowadzający mechanizmy ochronne, mające na celu zabezpieczenie autentyczności i integralności obiektów informacyjnych przed jakimikolwiek nieudokumentowanymi zmianami.

PDI jest więc zarówno pewnego rodzaju informatorem o pochodzeniu i historii obiektu informacyjnego, jego przynależności oraz powiązaniach z innymi obiektami w archiwum, jak i mechanizmem chroniącym jego integralność i autentyczność.

W celu powiązania obu komponentów pakietu informacyjnego model referencyjny OAIS przewiduje także element w postaci informacji o pakiecie (*Packaging Information*). Jego zadaniem jest identyfikacja poszczególnych składników pakietu informacyjnego.

Elementem niezbędnym w archiwum elektronicznym są wreszcie metadane przechowywanych obiektów (*Information Packages*). W modelu referencyjnym OAIS określane są one terminem *Descriptive Information*. Metadane dostarczają informacji o zawartości pakietu informacyjnego oraz umożliwiają jego odnalezienie w archiwum.

Pakiet informacyjny wraz ze wszystkimi jego elementami składowymi należy traktować jako obiekt archiwizacji w archiwum elektronicznym OAIS.

5.2. Identyfikacja obiektów sieciowych

W celu dotarcia do dokumentów opublikowanych w internecie najczęściej wykorzystuje się adresy URL (Uniform Resource Locators), które umożliwiają wyszukanie dokumentu oraz służą jako identyfikator w procesach cytowania i bibliograficznych odesłań do publikacji internetowych. Mogą być również stosowane w bazach danych, katalogach, indeksach, rejestrach i wszelkich innych typach bibliograficznych wykazów, odsyłających do pełnych tekstów dokumentów internetowych bądź ich metadanych. Jednak zmiana miejsca dokumentu sieciowego powoduje, że zastosowane odesłanie w postaci URL jest nieużyteczne, a więc obiekt cyfrowy przestaje spełniać podstawowe kryterium dostępności.

Należy więc zauważyć, że powszechnie stosowany URL nie powinien być określany mianem identyfikatora, lecz raczej „lokalizatora" obiektu sieciowego, ponieważ wskazuje jedynie lokalizację obiektu, a nie identyfikuje jednoznacznie samego obiektu.

Połowicznym rozwiązaniem jest stosowanie metod zapewniających tak zwaną stabilność okresową obiektów cyfrowych. Do metod tych zalicza się:

- zastosowanie systemu adresowania URL, w którym serwer dynamicznie ustala miejsce zapisu obiektu sieciowego, korzystając z odpowiednich skryptów oraz baz danych zawierających bieżącą lokalizację dokumentów,
- zastosowanie odpowiedniej konfiguracji serwera Web, która umożliwi przekierowanie z nieaktualnego do nowego adresu w formie tzw. *redirects* lub *aliases*,
- przeprowadzanie okresowej kontroli dostępności adresów i powiązanych z nimi obiektów (tzw. URL-Checks) przez administratora i wykonanie uaktualnienia odwołań do dokumentów.

Powyższa metodologia stanowi jednak rozwiązanie krótko- lub średniookresowe. Dzieje się tak z wielu powodów, głównie z racji bardzo prawdopodobnych zmian w metodologii adresowania, wynikających na przykład z technicznych modyfikacji otoczenia systemowego. Za sensowną uznaje się okresową kontrolę URL, jednak tylko przy założeniu tzw. „konsekwentnej pielęgnacji", która oznacza, że w przypadku stwierdzenia, że hiperłącze nie odsyła do pożądanego obiektu, należy ustalić źródło błędu, odszukać właściwy adres do obiektu i nanieść stosowne zmiany we wszelkich wykazach, katalogach, bibliografiach, portalach *etc.*, które do danego obiektu odsyłają. Są to zabiegi pracochłonne. Okresową niedostępność adresów URL mogą też

powodować błędy sieciowe lub niestabilne połączenia z serwerem. Wreszcie dokumenty sieciowe ulegają zmianom w wyniku procesów zachodzących w instytucjach, w których są zlokalizowane i ich identyfikacja oraz adresowanie za pomocą samego URL mogą okazać się zawodne.

W związku z powyższym zachodzi potrzeba zastosowania trwałego mechanizmu archiwizacji obiektów cyfrowych. Zaproponowane rozwiązanie to identyfikatory trwałe (ang. *persistent identifiers*) (PI).

5.3. Systemy trwałej identyfikacji obiektów sieciowych

Identyfikator trwały (PI) to niezmienna (określana też jako stabilna, unikatowa, permanentna) nazwa, którą przyporządkowuje się do obiektu sieciowego jeden raz na cały cykl jego „życia". Zadaniem PI jest jednoznaczna i trwała identyfikacja obiektu sieciowego oraz przynależnych do niego metadanych, niezależnie od miejsca (instytucji), w którym obiekt został zapisany i jest archiwizowany, z uwzględnieniem różnorodnych systemów, ich ograniczeń (granic), zmian oraz w obliczu występowania obiektów cyfrowych w różnych wersjach, postaciach, formach reprezentacji. Na podstawie PI, system obsługi PI powinien umożliwić zlokalizowanie dokumentu i jego odczyt. Obecnie wykorzystywane są głównie trzy systemy PI, tj. PURL, Handle System i URN. Bez względu na wybór zastosowanego systemu ważne jest, aby identyfikatory pozostawały niezmienne. Istotne jest także, aby dany system obsługi PI miał podbudowę instytucjonalną.

5.3.1. PURL – Persistent URL

System Persistent Uniform Resource Locator (PURL) jest rozwinięciem koncepcji URL i funkcjonalnie jest z nim tożsamy. System ten wykorzystuje adresy URL, które zamiast wskazywać na określony obiekt, wskazują na usługę przekierowującą do danego obiektu. Tak więc PURL składa się z adresu serwera usługi przekierowującej oraz identyfikatora obiektu, do którego chcemy uzyskać dostęp. Adresy PURL stosuje się wówczas, gdy przewiduje się częste zmiany położenia poszczególnych obiektów WWW. Pełnią one rolę oficjalnych adresów, pod którymi można znaleźć żądane zasoby, a odpowiednimi przekierowaniami zajmuje się serwer[9]. Baza danych serwera usługi

[9] A. Freedman, *Encyklopedia komputerów*, Gliwice 2004, s. 667.

przekierowań zawiera wszystkie identyfikatory zarejestrowane w danym systemie wraz z przypisanymi im aktualnymi lokalizacjami dokumentu. Można więc powiedzieć, że w systemie tym odróżnia się „identyfikatory" od „lokalizatorów", czyli adresów lokalizacji, w których przechowywane są kopie danego obiektu. W przypadku gdy mamy do czynienia z obiektem sieciowym, lokalizatory mają postać aktualnych URL poszczególnych kopii obiektu.

System PURL został wprowadzony przez Online Computer Library Center (OCLC) w 1995 roku, w ramach inicjatywy „Internet Cataloging Projects", której celem było poprawienie (dookreślenie, uściślenie) adresów internetowych zasobów, wykazywanych w katalogach bibliotecznych.

Składnia adresu PURL wygląda następująco: <Protocol><RA><Name>
Przy czym:

- Protocol to standardowy protokół, np.: http,
- RA to adres serwera usługi przekierowującej do wybranego obiektu,
- Name to nazwa wskazująca na określony obiekt.

Przykład: http://purl.oclc.org/keith/home,
gdzie:

- http – protokół,

purl.oclc.org – adres serwera przekierowującego,
/keith/home – nazwa zasobu.

System ten znalazł zastosowanie m.in. w Bibliotece Kongresu oraz United States Government Printing Office (GPO), eksperymentalnie również w OCLC. Aktualnie system PURL nie jest już rozwijany, natomiast zasady jego działania wykorzystano przy opracowywaniu bardziej kompleksowych systemów, takich jak Handle System i URN.

Najszerzej wykorzystywaną implementacją założeń systemu PURL jest Archival Resource Key – ARK[10]. Stanowi on schemat identyfikacyjny służący do trwałej dostępności cyfrowych obiektów. Identyfikator ARK jest stosowany jako link:

- odsyłający od obiektu cyfrowego do organizacji, do której obiekt przynależy,
- łączący obiekt cyfrowy z jego metadanymi,
- odsyłający do treści obiektu bądź jego kopii.

System ARK znalazł zastosowanie w 15 repozytoriach, m.in. w California Digital Library, Library of Congres, National Library of France.

[10] *Archival Resource Key*, [online:] http://www.cdlib.org/inside/diglib/ark/ [dostęp: 20.12.2008].

Trwałość w tym systemie identyfikacyjnym jest zapewniana przez usługodawcę, a nie składnię nazwy. ARK wskazuje metadane o obiekcie, nie daje gwarancji trwałości identyfikatora, zezwala na integrację innych schematów, a także jego zintegrowanie z innymi schematami.

Składnia ARK jest następująca: http://<NMAH>/ark:/<NAAN>/<Name>

Przy czym:

- NMAH to adres serwera usługi przekierowującej,
- NAAN to identyfikator instytucji nadającej poszczególnym obiektom identyfikatory we własnej przestrzeni nazw,
- Name to nazwa (identyfikator) przydzielona do danego zasobu.

Przykład: http://bnf.fr/ark:/13030/tf5p30086k

5.3.2. Handle-System

Handle-System[11] jest systemem identyfikatorów przypisywanych obiektom cyfrowym niezależnie od ich fizycznego umiejscowienia. Założenia systemu zostały opracowane przez Corporation for National Research Initiatives CNRI[12] i opisane w dokumencie RFC 3650[13].

W dokumencie tym autorzy zdefiniowali m.in. zasadę budowy identyfikatorów, na które składa się prefiks oraz sufiks. Prefiks jest numerycznym kodem, oznaczającym instytucję, która została zarejestrowana w Global Handle Service (instytucji nadzorującej system) jako upoważniona do nadawania obiektom identyfikatorów we własnej przestrzeni nazw. Sufiks identyfikatora jest nazwą (identyfikatorem) danego obiektu, unikatową w przestrzeni nazw danej instytucji i może składać się z dowolnej liczby znaków zgodnych z systemem ASCII.

Składnia Handle-System wygląda następująco: Handle: <HNA> / <HLN>, przy czym:

- HNA – prefiks instytucji nadawany przez Global Handle Service,
- HNL – identyfikator obiektu w przestrzeni nazw danej instytucji.

---

[11] *The Handle System*, [online:] http://www.handle.net/ [dostęp: 20.12.2008].

[12] CNRI – to amerykańska organizacja non profit, założona w 1986 roku, której głównym celem jest wspieranie rozwoju kluczowych technologii przetwarzania i udostępniania wiedzy z użyciem sieci komputerowych. Źródło: Corporation for National Research Initiatives: http://www.cnri.reston.va.us/about_cnri.html [dostęp: 20.12.2009].

[13] S. Sun, L. Lannom, B. Boesch, *Handle System Overview. Request for Comments*: *3650*, CNRI, November 2003, [online:] http://www.ietf.org/rfc/rfc3650.txt [dostęp: 20.12.2009].

Przy rejestracji dany obiekt otrzymuje identyfikator, do którego przypisane są informacje uzupełniające. Handle-System nie narzuca sztywnej struktury metadanych powiązanych z obiektem, więc zarówno rodzaj, jak i zakres tych informacji determinowany jest przez instytucję rejestrującą oraz typ obiektu cyfrowego. Wśród informacji o obiekcie najczęściej znajdują się dane właściciela (autora), opis dokumentu (tytuł, słowa kluczowe) oraz co najmniej jeden wpis pozwalający na dostęp do kopii danego obiektu.

Identyfikatory wraz z powiązanymi metadanymi przechowywane są w centralnej, ogólnodostępnej bazie danych, umożliwiającej szybkie uzyskanie podstawowych informacji na temat określonych obiektów poprzez usługi dostępne w sieciach komputerowych. Funkcje systemu umożliwiają jednostkom rejestrującym dystrybucję, administrację oraz rozwiązywanie (likwidację) identyfikatorów.

Z Handle-System korzysta obecnie wiele instytucji i firm. Przykładem zastosowania Handle-System są m.in. CORDA/ADL i DVIA, czyli systemy Departamentu Obrony Stanów Zjednoczonych, rejestrujące i zarządzające dokumentami związanymi z obronnością Stanów Zjednoczonych. Handle-System jest również użyteczny w projekcie DSpace realizowanym przez MIT, w którego ramach tworzona jest baza danych na temat materiałów edukacyjnych powstających we wszystkich wydziałach i jednostkach tej instytucji. Jeszcze innym projektem stosującym opisywany system jest The National Digital Library. Program, którego założeniem jest digitalizacja i utworzenie bazy danych dzieł zgromadzonych w bibliotekach publicznych i uczelnianych w Stanach Zjednoczonych[14].

Struktura identyfikatorów Handle-System pozwala także na rejestrację tzw. rejestratorów lokalnych, wówczas zarejestrowana instytucja ma możliwość rejestracji instytucji sobie podległych, które dysponują własną przestrzenią nazw dla obiektów cyfrowych. W takim przypadku prefiks identyfikatora składa się z dwóch numerycznych członów oddzielonych kropką (np. 10.1000), przy czym pierwszy człon określa instytucję nadrzędną zarejestrowaną przez Global Handle Service, natomiast drugi jest identyfikatorem lokalnego rejestratora. Drzewiasta struktura Handle-System pozwoliła na powstanie podsystemów identyfikacyjnych, z których najpopularniejszym jest Digital Object Identifier – DOI.

DOI to identyfikator dokumentu elektronicznego, który jest do niego na stałe przypisany i w odróżnieniu od identyfikatora URL nie zależy od fizycz-

[14] Na podstawie informacji dostępnych na stronach Wolnej Encyklopedii – Wikipedia: http://pl.wikipedia.org/wiki/Handle_System [dostęp: 10.01.2009].

nej lokalizacji dokumentu. Zgodnie z definicją proponowaną w *Encyklopedii komputerów*[15] Digital Object Identifier to rozwiązanie pozwalające na przydzielanie dokumentom, publikacjom i wszelkim innym zasobom, dostępnym w internecie, stałych, niezmiennych nazw zamiast adresów URL.

Podstawowym założeniem systemu DOI jest identyfikacja oraz wymiana obiektów cyfrowych. Trwają również prace nad organizacyjnymi oraz technicznymi rozwiązaniami, umożliwiającymi zarządzanie obiektami cyfrowymi oraz powiązanie producentów i dostawców obiektów z użytkownikami[16].

Zarządzaniem systemu zajmuje się Międzynarodowa Fundacja DOI (*International DOI Foundation IDF*), która jest organizacją non profit, finansującą się ze składek członkowskich oraz sprzedaży prefiksów i numerów DOI. Fundacja DOI sprawuje kontrolę nad instytucjami i firmami, które uzyskały prawo do pełnienia roli *DOI Registration Agency* (RA). Podstawowym zadaniem RA jest przydzielanie identyfikatorów wydawcom (Publisher ID) i zapewnienie im infrastruktury umożliwiającej tworzenie identyfikatorów obiektów (Item ID) oraz zarządzanie metadanymi przypisanymi identyfikatorom DOI. Od agencji RA oczekuje się promocji systemu DOI oraz współpracy na rzecz jego rozwoju.

Struktura DOI stanowi od roku 2001 standard ANSI/NISO (Z39.84), a jej komponenty są implementacją założeń Handle-System. System DOI składa się z następujących komponentów: metadane, DOI jako identyfikator trwały (PI) oraz techniczna implementacja Handle-System. Identyfikatory DOI zgodnie z założeniami Handle-System są ciągami znaków ASCII. Składają się przedrostka i końcówki.

Przykład: 10.1000/182,

przy czym:

- 10.1000 to przedrostek, w którym znaki 10 informują, że chodzi o identyfikator DOI,
- 1000 to numer przypisany przez IDF wydawcy (*Publisher ID*),
- natomiast sufiks 182 to końcówka, która jest przypisana do określonego dzieła (*Item ID*).

Publisher ID jest przypisywany wydawcom, którzy zdecydowali się zarejestrować i korzystać z systemu DOI przez agencję, która ma do tego prawo. Item ID jest nadawany przez samego wydawcę, który powinien zagwarantować, że ID będzie unikalne dla każdej wydanej przez niego publikacji. Item ID może, ale nie musi być, numerem katalogowym publikacji pochodzącym

[15] A. Freedman, *Encyklopedia komputerów…*, s. 143.

[16] *The DOI System*, [online:] http://www.doi.org/ [dostęp: 19.12.2008].

z innych systemów rejestrowania, np. ISBN, ISSN. Poprawny sposób podawania odnośników do źródeł wygląda następująco: doi: 10.1000/182.

System DOI jest stosowany m.in. w agencjach praw autorskich, wydawnictwach i bibliotekach. Typowym przykładem zastosowania DOI jest identyfikowanie elektronicznych wersji publikacji naukowych w repozytorium SpringerLink, przy czym identyfikator DOI może otrzymać artykuł, całe czasopismo naukowe, rozdział w książce, plik multimedialny, program komputerowy *etc*.

5.3.3. URN – Uniform Resource Name

Historia systemu URN rozpoczęła się w 1990 roku i ma związek z projektowaniem architektury World Wide Web (WWW). URN został wprowadzony jako ujednolicona forma oznaczania zasobów internetowych. Formy i kierunki rozwoju sieci internet są kontrolowane przez organizację Internet Assigned Numbers Authority (IANA). To właśnie IANA oraz ściśle związana z nią grupa robocza o nazwie Internet Engineering Task Force (IETF) stanowią siłę napędową w rozwoju internetu i *de facto* dyktują standardy, których najbardziej znaną postacią są publikacje pod tytułem Requests for Comments (RFCs). W dokumencie RFC 1737[17] z 1994 roku dość precyzyjnie określono wymagania dotyczące schematu URN, natomiast trzy lata później, w publikacji RFC 2141[18] z 1997 roku zostały wymienione cele rozwoju identyfikatorów trwałych PI.

System URN został świadomie pomyślany jako schemat otwarty, zdolny do integracji z systemami istniejącymi, na przykład z identyfikatorami ISBN albo URL. Od ponad 10 lat URN[19] funkcjonuje jako standard adresowania obiektów w instytucjach objętych obowiązkiem takiego identyfikowania zasobów, aby były one dostępne długotrwale oraz niezależnie od tego, w której instytucji są przechowywane.

System URN cieszy się dużą popularnością. Jest stosowany m.in. w narodowych bibliotekach takich krajów jak Finlandia, Holandia, Austria, Szwajcaria i Wielka Brytania. Istnieje także możliwość integracji identyfikatorów

[17] K. Sollins, L. Masinter, *Functional Requirements for Uniform Resource Names*, [online:] http://www.ietf.org/rfc/rfc1737.txt [dostęp: 20.12.2008].

[18] R. Moats, *URN Syntax. Request for Comments*: 2141, AT&T, May 1997, [online:] http://www.ietf.org/rfc/rfc2141.txt [dostęp: 20.12.2008].

[19] *Uniform Resource Names. A Progress Report*, „D-Lib Magazine", February 1996, [online:] http://www.dlib.org/dlib/february96/02arms.html [dostęp: 20.12.2008].

URN wraz z istniejącymi numerycznymi systemami identyfikacyjnymi dokumentów, np. ISAN, ISSN, ISBN.

Identyfikatory URN składają się z kilku hierarchicznie ułożonych elementów, tj. z *Namespace IDentifier NID* (tzw. identyfikatora przestrzeni nazw) oraz z podporządkowanych mu subelementów (*SNID, NSS*).

Składnia identyfikatora wygląda następująco: urn: <NID> [: SNID] : <NSS> przy czym:

- NID – identyfikator przestrzeni nazw,
- SNID – identyfikator podprzestrzeni nazw (jeśli występuje),
- NSS – unikalny dla danej podprzestrzeni identyfikator zasobu (łańcuch znaków).

Jedną z podprzestrzeni nazw systemu URN jest system NBN – National Bibliographic Number. Został on opracowany w celu wyszczególnienia w bibliografiach narodowych publikacji cyfrowych, na przykład czasopism elektronicznych, rozpraw doktorskich i habilitacyjnych, także innych publikacji, stanowiących narodowe dziedzictwo cyfrowe i podlegających obowiązkowi wieczystej archiwizacji. Koncepcja systemu NBN zrodziła się w ramach popularnych inicjatyw bibliotek narodowych, *Conference of Directors of National Libraries (CDNL)* oraz *Conference of European National Librarians (CENL)*.

NBN jest implementacją założeń systemu URN, w związku z czym składnia jego identyfikatorów wygląda następująco: urn: NBM : <ICC> [:SNS] NBNstring, przy czym:

- ICC to dwuliterowy kod kraju według ISO 3166,
- SNS to podprzestrzeń nazw,
- NBNstring to identyfikator w podanej przestrzeni nazw,

Przykład: urn:NBN:de:kobv:23-2312.

System NBN jest ogólnoświatowym systemem używanym wyłącznie w Bibliotekach Narodowych i wykorzystywanym do jednoznacznej, trwałej identyfikacji zarówno dokumentów cyfrowych, jak i fizycznych. Biblioteki Narodowe przyjmują na siebie obowiązek zarządzania przestrzeniami nazw w obrębie danego kraju.

Podsumowanie

Składowanie i archiwizacja zasobów nauki i kultury w sieci ma sens wówczas, gdy zasoby te w każdej chwili, obecnie i w najbardziej odległej przyszłości, mogą być udostępniane i użytkowane. Instytucje tworzące repozytoria

cyfrowych zasobów decydują się na rozmaite systemy ich trwałego identyfikowania. Zaleca się, aby w procesie decyzyjnym, dotyczącym wyboru systemu identyfikacji instytucje uwzględniły następujące kryteria:

- **Standaryzacja**. Instytucje powinny skłaniać się do stosowania systemów, które zostały zaakceptowane jako standard, najlepiej o światowym zasięgu.
- **Wymagania funkcjonalne**. Wybierane systemy identyfikacyjne powinny charakteryzować się trwałością, jednoznacznością, światowym zasięgiem, niezależnością od miejsca składowania. Identyfikatory trwałe powinny odsyłać równocześnie do wielu kopii jednego obiektu.
- **Elastyczność, skalowalność**. Stosowane systemy powinny być skalowalne oraz zdolne do rozszerzenia o nowe funkcje, bez zaburzenia ich zgodności z przyjętym standardem.
- **Niezależność technologiczna i kompatybilność**. Systemy identyfikacyjne powinny być generyczne, niezależne od protokołów i technologii, a także kompatybilne z funkcjonującymi instalacjami i usługami.
- **Instalacje, polecenia (rekomendacje)**. Przy wyborze systemu należy uwzględnić jego akceptację i popularność w skali międzynarodowej.
- **Koszty oraz trwałość**. Kryterium wyboru systemu powinny być koszty systemu (zarówno wstępne, jak i dalszego utrzymania) oraz jego niezawodność.

Opisane w tym rozdziale systemy trwałej identyfikacji obiektów sieciowych w zasadzie spełniają wszystkie z wymienionych kryteriów i są najczęściej implementowane w profesjonalnych repozytoriach. Należy jednak zaznaczyć, że obok nich istnieje także szereg innych, mniej popularnych rozwiązań: ERRoL – Extensible Repository Resources Lokator, GRI – Grid Resource Identifier, GUUID/UUID – Globally Unique Identifier/Universal Unique Identifier, InfoURI, NLA – National Library of Australian, LSID – Life Science Identifier, POI – PURL-Based Object Identifier, XRI – Extensible Resource Identifier.