

Koncepcja systemu trwałych identyfikatorów internetowych dla obiektów dziedzictwa kulturowego

Autor: Marcin Werla (mwerla@man.poznan.pl)

Raport powstał w okresie czerwiec - październik 2013 r. w ramach prac zespołu ekspertów NIMOZ.



Spis treści

1.	Oczekiwany zakres funkcjonalny.....	2
2.	Propozycja rozwiązania.....	3
2.1.	Struktura identyfikatora.....	3
2.2.	Zgodność identyfikatora z systemem URN.....	4
2.3.	Usługa zarządzania identyfikatorami.....	4
2.4.	Usługa rozpoznająca identyfikatory.....	6

1. Oczekiwany zakres funkcjonalny

Przez trwały identyfikator internetowy obiektu dziedzictwa kulturowego należy rozumieć unikalny w skali globalnej ciąg znaków, który:

1. Pozwala się odnieść jednoznacznie od określonego obiektu, tzn. można jednoznacznie określić, z którym obiektem związany jest identyfikator.
2. Pozwala się odnieść od określonego obiektu w sposób interaktywny, tzn. dysponując identyfikatorem można w stosunkowo prosty sposób przejść do dostępnej on-line autorytatywnej informacji na temat obiektu z którym związany jest identyfikator.
3. Jest trwały, tzn. prowadzi on do informacji na temat obiektu, z którym związany jest identyfikator, nawet jeżeli informacja ta zmienia swoją wirtualną lokalizację (adres URL).

W celu utworzenia systemu trwałych identyfikatorów internetowych niezbędne są:

- Stabilna instytucja zarządzająca o ugruntowanej pozycji, która na poziomie organizacyjnym umożliwi zapewnienie trwałości i wiarygodności systemu identyfikatorów i zgromadzonych w nim informacji.
- Sprawny system gromadzenia i aktualizacji danych powiązanych z identyfikatorami, na poziomie technicznym pozwalający na wysoką automatyzację działań związanych z utrzymaniem systemu identyfikatorów, przy jednoczesnym zapewnieniu jakości danych.
- Przyjazna usługa internetowa z prostym adresem WWW pozwalająca na rozpoznawanie identyfikatorów i dostarczająca użytkownikom powiązane z nimi informacje.

System trwałych identyfikatorów internetowych dla obiektów dziedzictwa kulturowego wykorzystywany będzie równolegle przez wiele instytucji posiadających tego typu zbiory, a koordynowany i utrzymywany powinien być centralnie przez jedną instytucję.

Z punktu widzenia instytucji posiadającej zbiory dziedzictwa kulturowego system trwałych identyfikatorów internetowych powinien umożliwiać wykonanie w sposób zupełnie zautomatyzowany następujących czynności:

- Zarejestrowanie obiektu - uzyskanie identyfikatora dla obiektu
- Przesłanie/aktualizacja danych dotyczących obiektu o określonym identyfikatorze (w szczególności wskazanie adresu WWW pod którym dostępne są cyfrowe wizerunki obiektu)
- Unieważnienie identyfikatora obiektu (sytuacja wyjątkowa)

Rozpoczęcie korzystania z takiej usługi powinno wiązać się z podpisaniem porozumienia czy umowy pomiędzy operatorem usługi, a instytucją korzystającą z usługi. Umowa taka powinna dawać instytucji swobodę zarządzania identyfikatorami i informacjami dotyczącymi jej własnych obiektów (w szczególności uzyskiwania identyfikatorów dla kolejnych obiektów oraz aktualizacji danych), przy jednoczesnym narzuceniu regulaminu wymagań technicznych i jakościowych, co do przesyłanych danych i zasad korzystania z serwisu.

Technicznie system powinien być skonstruowany tak, aby gwarantować spełnienie trzech cech podanych na początku niniejszego rozdziału dla każdego zarejestrowanego w systemie identyfikatora.

2. Propozycja rozwiązania

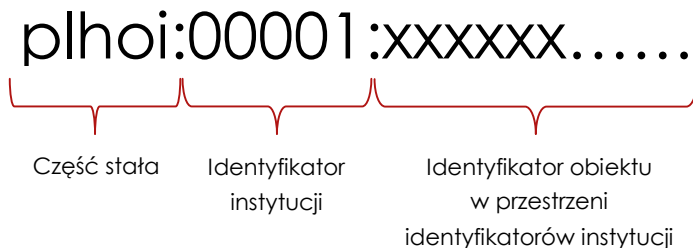
2.1. Struktura identyfikatora

Identyfikator powinien składać się z trzech następujących po sobie części:

1. **Pole o stałej wartości charakterystycznej dla tego typu identyfikatora**, informującej że dalszą część identyfikatora należy interpretować właśnie jako właściwą część informacyjną identyfikatora (tak jak powszechnie rozpoznawalny i łatwy w interpretacji jest identyfikator książki poprzedzony np. skrótem ISBN). Proponowana stała wartość to **'plhoi'** (**P**olish **H**eritage **I**dentifier) ułatwiająca rozszerzenie stosowania systemu identyfikatorów na inne sektory kultury.
2. **Pole zawierające wartość identyfikującą jednoznacznie konkretną instytucję, która rejestruje obiekty w systemie** (unikalna wartość dla każdej z instytucji) – dzięki takiemu polu jest pewność, że każdy identyfikator zarejestrowany w systemie jest unikalny, nawet jeżeli dwie różne instytucje będą chciały rejestrować taki sam identyfikator, gdyż zawsze różna będzie część składowa charakterystyczna dla danej instytucji. Proponowana jest **wartość numeryczna o stałej długości pięciu cyfr** (uzupełniana zerami, np. 00123), przydzielana poszczególnym instytucjom w kolejności ich rejestracji w systemie.
3. **Pole zawierające wartość unikalną dla przestrzeni identyfikatorów danej instytucji, identyfikującą konkretny obiekt**. Tymi wartościami powinna zarządzać dana instytucja, wskazane jest też ponowne użycie istniejących już identyfikatorów (numerów ewidencyjnych obiektów). Długość tego pola jest dowolna. Ze względów technicznych proponuje się, aby ograniczyć znaki dozwolone w tym polu do listy znaków, które mogą znaleźć się w adresie URL, z wyłączeniem znaku ~ (tylda). A więc byłyby to **duże i małe litery alfabetu łacińskiego** (bez liter diakrytyzowanych), **cyfry arabskie oraz znak kropki (.), minusa (-) i podkreślenia (_)**.

W celu zgodności z konwencją zapisu identyfikatorów URN¹ proponuje się, aby powyższe trzy części były rozdzielone znakiem dwukropka. Kompleksowa zgodność ze standardem URN jest opisana w następnym podpunkcie.

Zgodny z powyższymi założeniami przykładowy identyfikator mógłby zatem wyglądać następująco:



¹ Moats, Ryan (May 1997). "Request for Comments: 2141: URN Syntax". IETF. <http://tools.ietf.org/html/rfc2141>

2.2. Zgodność identyfikatora z systemem URN

Identyfikator o opisanej powyżej strukturze może być w prosty sposób doprowadzony do postaci zgodnej ze specyfikacją URN. W celu zgodności składniowej należy po prostu dodać do identyfikatora prefiks „urn:”. W celu zgłoszenia identyfikatora typu „plhoi” jako oficjalnej formalnej przestrzeni nazw URN² należałoby zarejestrować go w organizacji IANA (<http://www.iana.org/>) zgodnie ze stosowną procedurą³.

Warto tutaj odnotować, że fundacja zarządzająca powszechnie stosowanymi identyfikatorami publikacji elektronicznych DOI od dłuższego czasu wstrzymuje się z rejestracją tego identyfikatora jako oficjalnej przestrzeni nazw URN, podając jako uzasadnienie brak praktycznych korzyści z takiej rejestracji. Fundacja DOI utrzymuje, że gdy wdrożony zostanie powszechnie dostępny i ogólny system rozpoznawania identyfikatorów URN i przekształcania ich na lokalizacje sieciowe, wtedy DOI zostanie poddane standaryzacji URN⁴.

Praktycznym podejściem może być podjęcie próby rejestracji „plhoi” jako oficjalnej formalnej przestrzeni nazw URN i wstrzymanie działań, jeżeli okaże się że procedura ta przedłuża się, jest wyjątkowo pracochłonna lub kosztowna.

2.3. Usługa zarządzania identyfikatorami

Institucje posiadające zbiory dziedzictwa kulturowego i korzystające z usługi trwałych identyfikatorów internetowych powinny mieć autoryzowany dostęp do usługi pozwalającej na zarejestrowanie nowego obiektu i uzyskanie dla niego trwałego identyfikatora, aktualizację danych dotyczących obiektu, oraz (w wyjątkowych sytuacjach) unieważnienie identyfikatora obiektu. Sytuacje takie jak zniszczenie obiektu fizycznego nie powinny się wiązać z unieważnieniem identyfikatora obiektu. Zamiast tego należy uzupełnić w systemie informacje o obiekcie, odnotowując fakt jego zniszczenia. W przypadku unieważnienia identyfikatora obiektu, identyfikator ten nie powinien być już nigdy ponownie wykorzystywany w przyszłości. Ponadto wszystkie zmiany wprowadzane do systemu powinny być odnotowywane, tak aby dało się prześledzić pełną historię zmian informacji skojarzonych z danym obiektem, wraz z informacją kiedy dokładnie i przez kogo te zmiany były wprowadzane.

Powyższa funkcjonalność powinna być dostępna w dwojaki sposób:

1. Poprzez aplikację internetową, umożliwiającą autoryzowanemu pracownikowi instytucji stosowne działania przy pomocy przeglądarki WWW.
2. Poprzez usługę sieciową (zalecana jest usługa typu REST), umożliwiającą integrację usługi zarządzania identyfikatorami z systemami informatycznymi poszczególnych instytucji.

Dostęp do usługi zarządzania identyfikatorami powinien być autoryzowany. Rozpoczęcie korzystania z usługi przez daną instytucję powinno wiązać się z podpisaniem umowy regulującej zasady współpracy, a w szczególności podkreślającej obowiązek bieżącego aktualizowania danych przekazanych do systemu. Jednym z efektów podpisania umowy powinno być przydzielenie instytucji unikalnego pięciocyfrowego identyfikatora (por. punkt 2.1. Struktura identyfikatora). Ponadto powinien być wtedy utworzony w systemie użytkownik z uprawnieniami administracyjnymi dla danej instytucji. Użytkownik taki powinien móc tworzyć kolejnych użytkowników na potrzeby danej instytucji wg modelu uprawnień przedstawionego poniżej.

² Official IANA Registry of URN Namespaces <http://www.iana.org/assignments/urn-namespaces/>

³ Daigle, Leslie L.; van Gulik, Dirk-Willem; Falstrom, Patrik (October 2002). "Request for Comments: 3406: Uniform Resource Names (URN) Namespace Definition Mechanisms". IETF. <http://tools.ietf.org/html/rfc3406>

⁴ "Factsheet: DOI System and Internet Identifier Specifications". International DOI Foundation. October 2012. <http://www.doi.org/factsheets/DOIIdentifierSpecs.html>

Poziom uprawnień	Zakres operacji
Anonimowy	Odczyt aktualnej wersji danych dostępnych w systemie (bez dostępu do historii zmian) dla dowolnego obiektu.
Ograniczony	Odczyt wszystkich danych (również historii zmian) wprowadzonych do systemu przez wszystkich użytkowników danej instytucji.
Podstawowy	Jak poziom „ograniczony”, a ponadto możliwość dodawania nowych obiektów do systemu oraz możliwość edycji (w tym usuwania) obiektów wprowadzonych przez tego użytkownika.
Rozszerzony	Jak poziom „ograniczony”, a ponadto możliwość dodawania nowych obiektów do systemu oraz możliwość edycji (w tym usuwania) obiektów wprowadzonych przez dowolnego użytkownika w ramach danej instytucji.
Administracyjny	Jak poziom „rozszerzony”, a ponadto możliwość zarządzania użytkownikami w ramach danej instytucji.
Super-administracyjny	Dostęp do wszystkich danych w systemie, możliwość edycji wszystkich danych w systemie ⁵ , możliwość zarządzania wszystkimi użytkownikami, możliwość rejestrowania nowych instytucji. Ten poziom uprawnień powinien być zastrzeżony wyłącznie dla pracowników odpowiedzialnych za zarządzanie całym systemem.

Przedstawiony powyżej model nie zakłada konieczności ograniczania odczytu aktualnej wersji informacji zebranych w systemie.

Tak więc autoryzowany użytkownik z odpowiednim poziomem uprawnień powinien być w stanie wykonać operacje takie jak:

- Przeszukiwanie i przeglądanie identyfikatorów, opcjonalnie z zawężeniem do identyfikatorów wprowadzonych w ramach kont użytkowników należących do danej instytucji.
- Odczyt aktualnej wersji danych powiązanych z określonym identyfikatorem.
- Odczyt historii zmian danych powiązanych z określonym identyfikatorem.
- Dodanie nowego identyfikatora i powiązanych z nim danych.
- Aktualizacja danych powiązanych z określonym identyfikatorem.
- Usunięcie danych powiązanych z określonym identyfikatorem.

Operacja wymagające podania konkretnego identyfikatora (odczyt, zapis, modyfikacja) powinny być również możliwe do wykonania w trybie wsadowym, tzn. operującym na wielu obiektach jednocześnie. W praktyce może być to zrealizowane np. poprzez umożliwienie użytkownikom zaimportowania do systemu arkusza kalkulacyjnego o określonej strukturze z nowymi lub zmienionymi danymi.

Zakres danych skojarzony z pojedynczym identyfikatorem może być różny, w zależności od zasad współpracy z instytucjami dostarczającymi dane oraz integracji z usługą agregującą metadane obiektów. Przyjąc można następujące możliwości:

- Brak dodatkowych informacji – system wie jedynie, że identyfikator jest zarejestrowany w ramach danej instytucji.
- Adres WWW prowadzący do strony danej instytucji zawierającej szczegółowe informacje o obiekcie lub ewentualnie adres WWW prowadzący bezpośrednio do cyfrowego wizerunku obiektu – np. do pliku JPG.

⁵ Zasady modyfikacji danych przez operatora usługi powinny być określone w umowie pomiędzy dostawcą danych a operatorem i każdorazowo konsultowane z dostawcą danych.

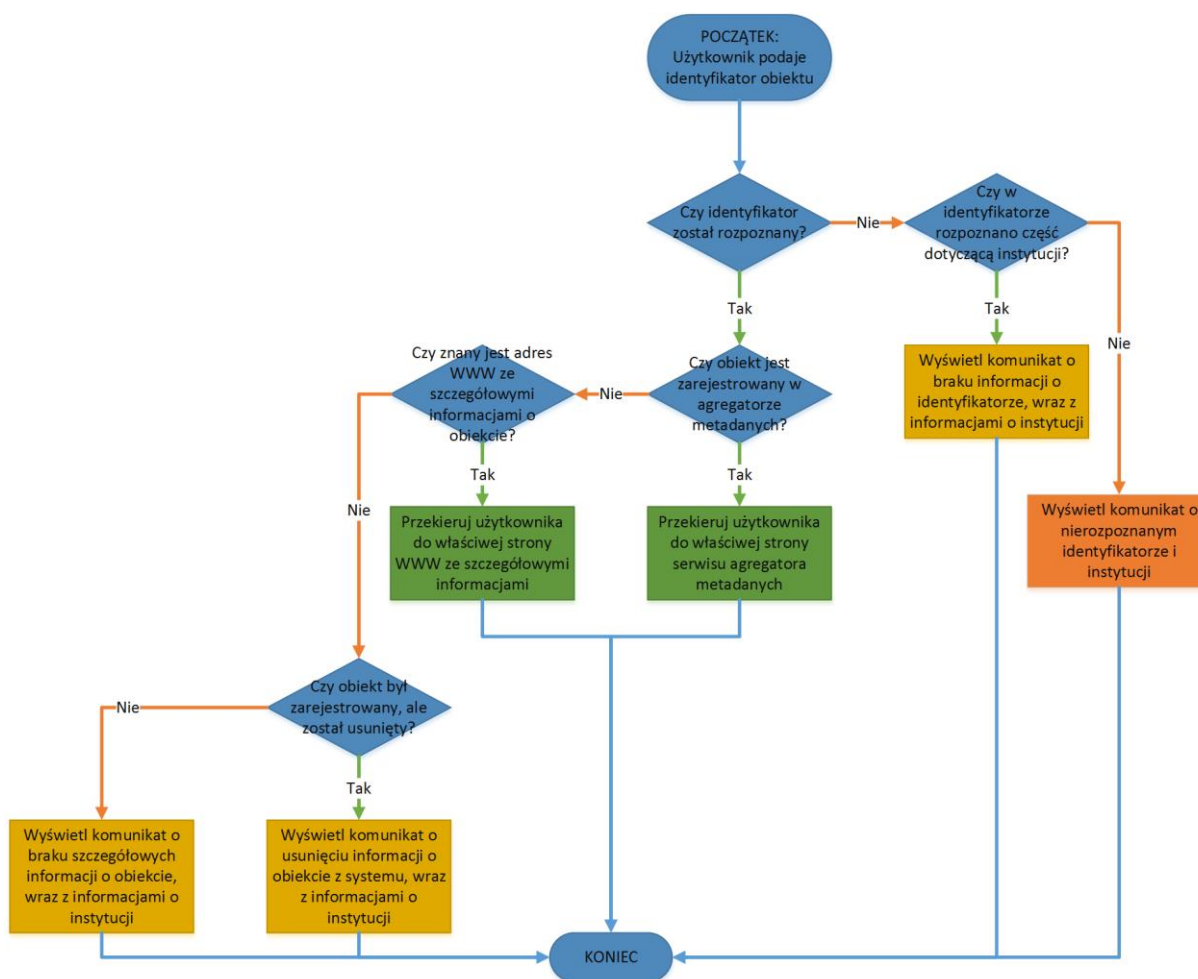
- Powiązanie z danymi zebranymi na temat obiektu w usłudze agregującej metadane.

Scenariusze zachowania systemu w powyższych przypadkach opisano w następnym punkcie.

2.4. Usługa rozpoznająca identyfikatory

Aby korzystanie z trwałych identyfikatorów internetowych było wygodne w praktyce, niezbędne jest udostępnienie usługi rozpoznającej identyfikatory. Usługa ta powinna być dostępna pod prostym i krótkim adresem. Ze względu na podniesienie wiarygodności adresu warto rozważyć uruchomienie takiego serwisu w domenie instytucji zarządzającej np. <http://id.nimoz.pl/>. Wejście na główną stronę serwisu powinno prezentować użytkownikowi prosty formularz do wprowadzenia identyfikatora do rozpoznania⁶ oraz odpowiednie informacje wprowadzające.

Rysunek poniżej prezentuje możliwy algorytm przetwarzania żądania rozpoznania identyfikatora, w zależności od danych jakie są dostępne na temat danego obiektu w systemie.



⁶ Por. Analogiczny formularz zbliżonej usługi funkcjonujące w ramach Federacji Bibliotek Cyfrowych (<http://fbc.pionier.net.pl/id>).

Przyjęto tu założenie, że poza systemem rozpoznawania identyfikatorów, który gromadzi głównie adresy URL dla poszczególnych obiektów, istnieje również agregator metadanych gromadzący bardziej szczegółowe informacje, i założono że agregator ten jest preferowany w stosunku do odnośników bezpośrednio do stron WWW dostawcy danych.

Użycie usługi powinno być możliwe poprzez wprowadzenie identyfikatora obiektu do wspomnianego powyżej formularza i wysłanie go. Alternatywnie powinno być też możliwe wywołanie usługi poprzez wprowadzenie adresu URL w postaci <główny adres usługi rozpoznającej>/<identyfikator obiektu>, np.: <http://id.nimoz.pl/plmoi:00026:M.Ob.953>. Niezależnie od sposobu wywołania (pośrednio przez formularz czy bezpośrednio przez identyfikator w adresie URL), zachowanie usługi powinno być takie samo. Zaleca się też stosowanie kodów odpowiedzi HTTP zgodnych typami odpowiedzi PURL⁷.

⁷ Por. Dokumentacja serwisu PURL utrzymywanego przez OCLC (<http://purl.oclc.org/docs/help.html#overview>).