21

Lucjan Stalmach *
Biblioteka Medyczna Collegium Medicum UJ
Ośrodek Komputerowy CM UJ
e-mail: lucek@bm.cm-uj.krakow.pl

NOWE MECHANIZMY DYSTRYBUCJI INFORMACJI W INTERNECIE. METADANE I ICH STANDARYZACJA

[NEW METHODS OF INFORMATION DISTRIBUTION ON THE INTERNET. METADATA AND THEIR STANDARDIZATION]

Abstrakt: Cechą charakterystyczną Internetu jest ogromna zawartość danych w nim dostępnych oraz ich duża zmienność. Efektywne wykorzystanie tych danych wymaga stworzenia nowych mechanizmów przetwarzania, które dostosowane będą do różnorodności informacji pod względem tematyki, formatu, miejsca i czasu jej publikacji. Nowe możliwości w tym zakresie stwarza coraz szersze wykorzystanie standaryzowanego zapisu metadanych, jako formy opisu informacji. Omówiono standardy związane z wykorzystaniem metadanych, takie jak: Dublin Core, RDF, RSS, XML, OpenURL oraz związki zachodzące między nimi.

INTERNET - METADANE - STANDARDY

Abstract: What is characteristic for the Internet is a huge number of data accessible in it and their great changebility. To utilize these data efficiently we should create new mechanisms for their processing, which will be adequate to the great variability of their subject, formats, place and time of publication. Standardized recording of metadata seen as forms of information description creates new possibilities in this field. The author discuss briefly the standards connected with the use of metadata, such as: Dublin Core, RDF, RSS, XML, OpenURL and relations among these standards.

INTERNET - METADATA - STANDARDS

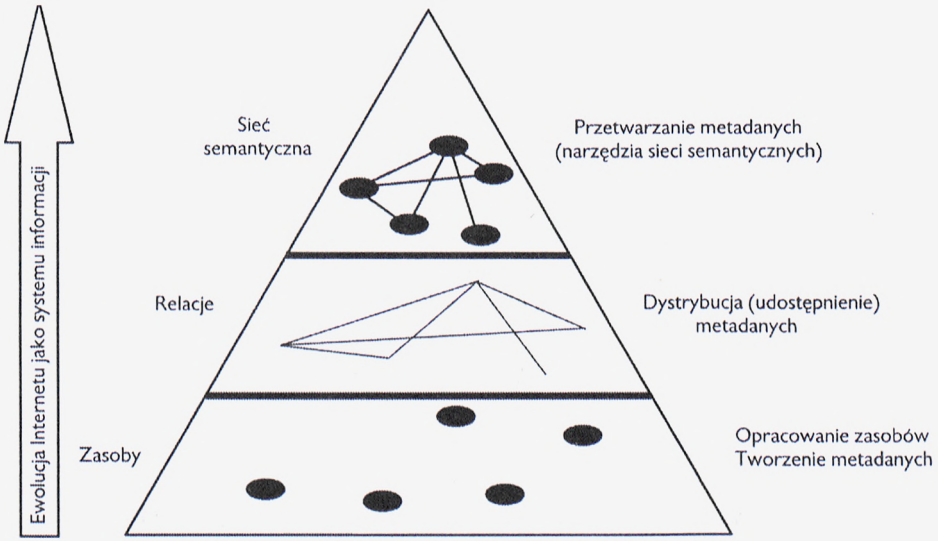
* Mgr inż. LUCJAN STALMACH, informatyk, kierownik Działu Technologii Informatycznych w Bibliotece Medycznej Collegium Medicum UJ. Zajmuje się wdrażaniem narzędzi informatycznych w obszarze przetwarzania informacji i wspomagania procesów bibliotecznych. Jest współautorem systemu dystrybucji dokumentów elektronicznych, autorem systemu metawyszukiwania w medycznych bazach bibliograficznych, programu do analizy publikacji pracowników CM UJ w bazie PubMcd. Ostatnie publikacje: (2002) System dystrybucji skanowanych zasobów bibliotecznych w Collegium Medicum Uniwersytetu Jagiellońskiego; (2001) Realizacja systemu metawyszukiwania informacji w medycznych bazach internetowych.

*
*

WPROWADZENIE

W pierwszych latach intensywnego rozwoju Internetu, w szczególności sieci Web, istotne było przede wszystkim to, aby informacja stała się dostępna w sieci i wzbogaciła dostępne w niej zasoby. Najczęstszą metodą wyszukiwania informacji było korzystanie z przeszukiwania oferowanego przez serwisy indeksujące, takie jak Altavista, Google, Excite. Każdy z tego typu serwisów indeksował jedynie pewien obszar informacji dostępnej w Internecie, co z kolei stworzyło potrzebę budowy systemów metawyszukiwania (np. Metacrowler, Vivísimo, Dogpile) integrujących wyniki pochodzące z wielu wyszukiwarek.

Ogromny przyrost informacji publikowanej w Internecie (2,8 min stron WWW w 1998 r. wobec 9 min w 2002 r., Web Characterization - OCLC http://wcp.oclc.org) oraz dynamika jej zmian sprawiają, iż dostęp do informacji jedynie za pomocą indeksujących serwisów wyszukiwawczych staje się niewystarczający. Konieczne jest opracowywanie nowych narzędzi dystrybucji informacji i narzędzi do jej przetwarzania.



Rys. 1. Mechanizmy tworzenia sieci semantycznych

Przemiany zachodzące w Internecie jako systemie informacji przyrównać można do ewolucji świata organicznego. Na wyższych poziomach ewolucji organizmy zaczynają tworzyć bardziej skomplikowane modele funkcjonowania, łącząc się w kolonie, stada, społeczności. Wytwarzają więzi (relacje) między sobą, definiują zasady współpracy i język porozumiewania się.

Podobnie dzieje się z Internetem jako ogromnym zbiorem pojedynczych źródeł informacji. Źródła te, wytwarzając między sobą relacje (np. poprzez cytowania, odnośniki), zaczynają tworzyć bardziej złożony system-sieć semantyczną (semantic Web), 'firn Barners--Lce określił tym pojęciem sieć Web, która wykazuje pewne cechy inteligencji, tzn. dostarcza na podstawie zaawansowanych mechanizmów filtrowania tylko te usługi i informacje, które zgodne są z wybranym przez użytkownika kontekstem znaczeniowym.

METADANE, SIECI SEMANTYCZNE

Aby pojedynczy obiekt mógł zaistnieć jako część składowa większego systemu, musi zidentyfikować się w tym systemie - musi mieć określoną pozycję w hierarchii, zadania do wykonania. W obszarze informacji tym, co identyfikuje zasoby, są metadane, czyli „dane o danych". Zdefiniowanie metadanych dla zasobu niewiele jednak wniesie do całego systemu informacji, jeśli nie będą one mogły być rozpoznane (przetworzone) przez system przy użyciu określonych narzędzi. Dopiero wówczas metadane nabiorąznaczenia-stanąsię charakterystyką semantyczną zasobu.

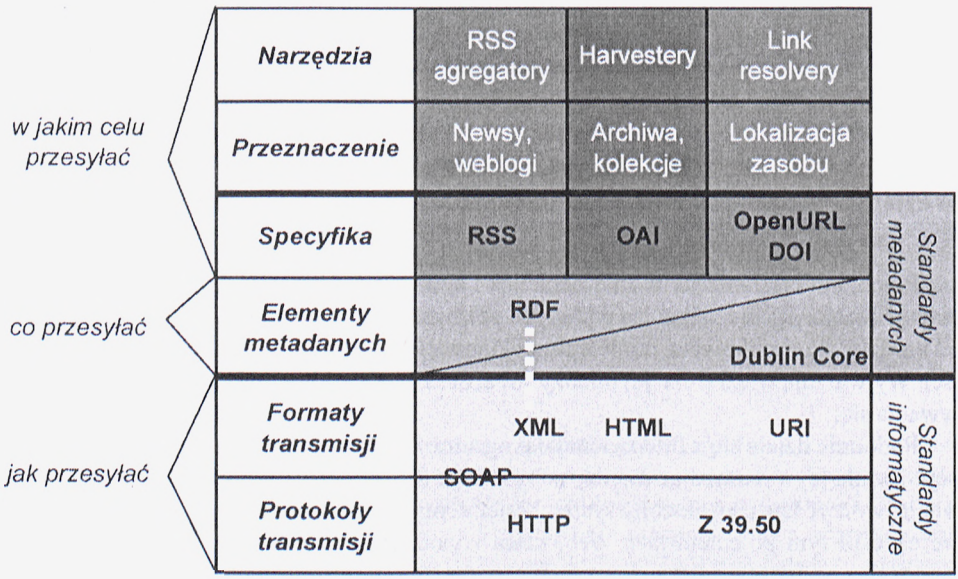
STANDARDY JAKO ELEMENT BUDOWY SIECI SEMANTYCZNYCH

Znaczenie, jakie odgrywają standardy dla metadanych można przyrównać do znaczenia wspólnego języka w komunikacji społecznej. To właśnie istnienie standardów kodowania i przesyłania umożliwia przetwarzanie metadanych i budowania na ich podstawie zaawansowanych usług informacyjnych.

W ostatnim okresie uczyniono wielki wysiłek w kierunku ujednolicenia reprezentacji metadanych w Internecie. Przyniosło to rezultat w postaci opracowania kilku standardów, które dla postronnego użytkownika wydawać się mogą konkurującymi ze sobą. Bliższa analiza, dokonana w tym artykule, pokazuje jednak, iż standardy te obejmują różne obszary zastosowań i wzajemnie dopełniają się lub z siebie wynikają.

PORZĄDEK W CHAOSIE, CZYLI „DZIEL I RZĄDŹ"

Aby wprowadzić porządek do licznego zbioru standardów, jakie dla potrzeb wymiany informacji w Internecie zostały opracowane, należy odnieść się przede wszystkim do problemów, jakie dany standard stara się uporządkować.



Rys. 2. Systematyka standardów związanych z transmisją metadanych w Internecie

Rozważając proces wymiany informacji, można w nim wyróżnić między innymi następujące problemy:

- jak przesłać informację,
- jaką informację przesłać,
- do jakich celów może być ona wykorzystana (po co jąprzesyłać, jakimi narzędziami ją przetwarzać).

Na rys. 2 przedstawiono systematykę standardów nawiązującą do wspomnianych wyżej problemów. Oczywiście dokonanie ścisłego przyporządkowania danego standardu do danej warstwy jest pewnym uproszczeniem, jednak pozwala określić wzajemne zależności między poszczególnymi standardami.

JAK PRZESYŁAĆ METADANE?

Pytanie o to, jak przesyłać metadane, zawiera w sobie zarówno kwestię ich transmisji, jak i ich formatu. Protokół transmisji określa reguły komunikacji między nadawcą i odbiorcą: sposób inicjacji komunikacji, potwierdzania odbioru danych, kończenia transmisji, zachowania w sytuacji wystąpienia błędów.

Z punktu widzenia sieci komputerowej metadane niczym nie różnią się od zwykłych danych przesyłanych w tych sieciach. Zatem protokoły transmisji są takie same. Najpopularniejszym obecnie protokołem transmisji jest protokół **HTTP** (Hypertext Transfer Protocol), kojarzony powszechnie z przeglądaniem dokumentów (stron WWW) w formacie HTML. Protokół ten wykorzystywany jest także do przeglądania (przesyłania) dokumentów w formacie XML oraz do tzw. tunelowania innych protokołów.

Powszechnie akceptowanym w środowisku bibliotek protokołem komunikacji między rozproszonymi katalogami OPAC i bazami bibliograficznymi jest protokół **Z39.50.** Protokół ten opracowany został przede wszystkim w celu standaryzacji rozproszonego przeszukiwania rozproszonych zasobów. Ponieważ nie jest to protokół ukierunkowany na sieć Web, należy przypuszczać, iż w przyszłości wyparty zostanie przez protokół HTTP lub SOAP. Protokół **SOAP** (Simple Object Access Protocol) jest rozwinięciem protokołu HTTP przeznaczonym do budowy mechanizmów komunikacji między aplikacjami internetowymi w ramach formatu XML.

Z protokołem transmisji wiąże się także format, w jakim dane będą przesyłane. Chodzi tu zarówno o strukturę tych danych, jak i na przykład o kodowanie znaków diakrytycznych. Dominujące obecnie formaty transmisji to **HTML** i **XML.** Formaty te określają składnię dokumentów oraz sposób ich kodowania. XML, w przeciwieństwie do HTML, pozwala uwzględnić semantykę dokumentu poprzez określenie relacji hierarchii poszczególnych elementów dokumentu.

Standaryzacja formatu danych objęła również standaryzację zapisu adresów rozproszonych zasobów. Wraz z formatem HTML i XML używany jest standard **URI** (Uniform Resource Identifier) określający sposób identyfikacji zasobu w sieci WWW. Najczęściej spotykanym typem URI jest adres **URL.**

CO PRZESYŁAĆ?

Opowiedź na pytanie,, jakie metadane transmitować?", nigdy nie będzie jednoznaczna, gdyż zależy od potrzeb użytkowników informacji. Niemniej jednak podjęto próbę wyróżnienia pewnego wspólnego dla wielu zastosowań zakresu metadanych.

Dublin Core Metadata Initiative (ISO 15836) definiuje 15 podstawowych, opcjonalnych elementów metadanych do opisu zasobów z różnych dziedzin i różnych typów. Część tych elementów odnosi się do zawartości zasobu, część do praw autorskich, a część charakteryzuje formę, w której zasób występuje. W praktyce, wybrane elementy DC kodowane są często w formacie XML lub, jak to się dzieje w przypadku stron WWW, w HTML jako znaczniki META.

Określenie wspólnego minimalnego zbioru elementów (DC element set) metadanych oczywiście nie do końca stanowi odpowiedź na pytanie „co przesyłać?". Różnorodność zasobów elektronicznych często sprawia, że konieczne jest scharakteryzowanie zasobu poprzez inne elementy metadanych. Tutaj użytkownikom przychodzi z pomocą standard definiowania metedanych **RDF** (Resource Description Framework). RDF sama w sobie niczego nie opisuje. Stanowi jednak elastyczny mechanizm (ramy) dla definiowania opisów wszelkiego rodzaju zasobów -- standard ten traktowany jest jako generyczny format metadanych. RDF ściśle nawiązuje do terminologii języka XML, który określa format dla metadanych zgodnych ze standardem RDF.

W JAKIM CELU PRZESYŁAĆ METADANE?

Najogólniej mówiąc, zarówno tworzenie, jak i przesyłanie metadanych ma na celu zaspokojenie potrzeb informacyjnych użytkowników. Nastąpi to wówczas, gdy użytkownik dostanie pełną, adekwatną, aktualną i wiarygodną informację.

Nie jest to zadanie łatwe, biorąc pod uwagę zmiany, jakim podlega informacja w Internecie. Jakże często zmienia się miejsce udostępniania informacji - nowe strony WWW pojawiają się i znikają, odnośniki dezaktualizują się. Coraz więcej informacji ma charakter czasowy. Warto zauważyć, iż coraz częściej Internet staje się dla agencji informacyjnych takim samym medium jak telewizja czy radio. Wszystko to prowadzi do sytuacji, gdzie niemożność dotarcia do informacji nie wynika z jej braku, ale z braku mechanizmów jej szybkiego odszukania. Fundamentalną rolę w tworzeniu takich mechanizmów odgrywają kolejne trzy standardy opracowane dla metadanych: RSS, OAI-PMH oraz OpenURL.

INFORMACJA O NOWOŚCIACH, NEWS, WEBLOGI

Zapewne każdy użytkownik Internetu posiada swoje ulubione serwisy (portale), których zawartość stara się śledzić na bieżąco. Co jednak zrobić w przypadku, gdy takich ulubionych źródeł informacji jest kilkadziesiąt lub kilkaset? Jak zorganizować ich przeglądanie, aby na bieżąco śledzić wszelkie nowości?

Dla tych właśnie celów opracowano standard **RSS** (Rich Site Summary, RDF Site Summaries lub Really Simple Syndication). Standard ten definiuje metodę rozgłaszania *(syndication)* informacji o nowościach lub zmianach w zawartości portalu i zbierania materiałów w celu udostępnienia ich użytkownikowi w sposób usystematyzowany chronologicznie i tematycznie przez tzw. kanały informacyjne.

Rozgłaszanie odbywa się przez publikowanie zbioru w formacie XML, w którym znajduje się opis nowych zasobów, newsów, blogów. Opis ten zgodny jest ze specyfikacją RDF.

Sprawne pobieranie zbiorów RSS (czyli informacji o tym, co nowego pojawiło się w portalu) oraz ich przetwarzanie zapewniają RSS agregatory, programy, które pozwalają zdefiniować użytkownikowi, z jakich portali mają być pobierane zbiory RSS informujące o nowościach.

ARCHIWA I KOLEKCJE ROZPROSZONE

Przetwarzanie rozproszonych zasobów opiera się na jednym z. dwu schematów:

• rozproszonym przeszukiwaniu. Przeszukiwanie rozproszonych kolekcji odbywa się synchronicznie (równocześnie), po czym wyniki są kompilowane,

• zbieraniu *(harvesting)* i przeszukiwaniu lokalnym. Informacja o zasobach w wielu rozproszonych kolekcjach zbierana jest w jednym systemie oferującym przeszukiwanie (już nierozproszone) pobranych informacji (metadanych).

Ze względu na to, iż rozproszone przeszukiwanie nie sprawdza się przy dużej ilości rozproszonych kolekcji, coraz popularniejszy staje się drugi z przedstawionych schematów.

Aby zapewnić zgodność stosowanych rozwiązań opracowano standard **OAI-PMH** (Open Archive - Protocol for Metadata Harvesting). Standard definiuje dwie grupy uczestników wymiany rekordów metadanych - dostawcę danych, który dostarcza na żądanie metadanych o swych lokalnych zasobach, oraz dostawcę usług, gromadzącego te dane (narzędziami zwanymi harvesterami) i oferującego użytkownikom końcowym możliwość ich przeszukiwania.

Protokół OAI-PMH wykorzystuje do transmisji protokół HTTP, a przesyłane dane kodowane są w formacie XML. Jako zakres wymienianych między kolekcjami metadanych przyjęto elementy zdefiniowane przez standard Dublin Core.

Przykładem rzeczywistego wykorzystania OAI-PMH jest projekt NDLTD (Networked Digital Library of Thesis and Dissertation), w którym ponad 200 instytucji buduje wspólny katalog prac doktorskich i magisterskich dostępnych w wersji elektronicznej.

LOKALIZACJA ZASOBÓW

Posiadanie informacji o zasobie nie jest równoznaczne z dostępem do tego zasobu. Dla przykładu, proces pozyskania artykułu pełnotekstowego z serwisu WWW wydawcy, na podstawie informacji bibliograficznej zawartej w bazie innego dostawcy, wiąże się z koniecznością integracji tych dwóch usług. Dodatkowo problem komplikuje kwestia identyfikacji użytkownika oraz jego uprawnienia do korzystania z zasobu.

Często bywa tak, iż poszukiwany artykuł nic jest dostępny dla użytkownika w jednej usłudze, ale dostępny jest w innej. Aby móc osiągnąć integrację dostępu do zasobów elektronicznych wielu dostawców oferuje wsparcie dla standardu **OpenURL,** definiującego składnię odnośników URL (linków) do zasobów oraz mechanizm ich przetwarzania.

Istotne jest, iż zdefiniowany mechanizm pozwala uwzględnić kontekst użytkownika- to, co użytkownik uzyska, odwołując się do danego odnośnika, zależne jest od tego, z jakich usług ma prawo korzystać.

PODSUMOWANIE

Krótka analiza przedstawionych standardów wykazuje, iż choć działają one w różnych warstwach i są przeznaczone do różnych zastosowań - wzajemnie się uzupełniają. Porządkują one różnorodne działania zmierzające do lepszego zaspokojenia potrzeb informacyjnych nie tylko użytkowników bibliotek, ale wszystkich użytkowników Internetu. Wydaje się oczywiste, iż to właśnie standardy pozwolą budować nowe narzędzia przetwarzania informacji i to one stanowić będą fundament dla budowy „inteligentnej sieci Web".

WYKORZYSTANE ŹRÓDŁA I OPRACOWANIA

Allen, J. [dok. elektr.]. Making a Semantic Web, http://www.nctcruciblc.com/semantic.litml [odczyt: 21.05.2004],

Apps, A. [dok. elektr.]. The OpenURL and OpenURL Framework: Demystifying Link Resolution, http://www.ariadnc. ac.uk/issuc38/apps-rpt/ [odczyt: 21.05.2004],

Brand, A.; F. Daly; B. Meyest (2003). Metadata demystified. Sheridan Press.

Louis, T. [dok. elektr.]. What is RSS, http://www.informit.com/articles/article.asp?p= 169476 [odczyt: 21.05.2004],

Powell, A. [dok. elektr.]. Five steps guide to becoming a content provider in the JISC Information Environment, http://www.ariadne.ac.uk/issuc33 [odczyt: 21.05.2004],