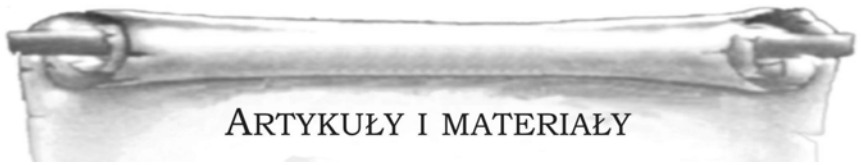
Marek Nahotko

Automatyczne tworzenie metadanych

Bibliotheca Nostra : śląski kwartalnik naukowy 2/2, 13-31

2010



MAREK NAHOTKO
Instytut Informacji Naukowej i Bibliotekoznawstwa
Uniwersytetu Jagiellońskiego

AUTOMATYCZNE TWORZENIE METADANYCH

Często powtarzane są informacje o szybkim tempie rozwoju zasobów Webu, porównywalnym do eksplozji[1]. Zasadnicze znaczenie dla organizacji i udostępniania tych zasobów, zróżnicowanych pod względem treści, jakości, formy i formatu, ma tworzenie metadanych o odpowiedniej jakości. Wszystkie najnowsze kierunki rozwoju zasobów sieciowych, takie jak powstawanie globalnej biblioteki cyfrowej (Nahotko 2010), rozumianej jako scentralizowane usługi wyszukiwania informacji w rozproszonych zasobach informacyjnych oraz ich ewolucja w kierunku usług Semantycznego Webu, wymagają efektywnego tworzenia metadanych.

W obecnym środowisku sieciowym metadane są tworzone zarówno w procesach realizowanych przez ludzi, jak i bez ich bezpośredniego udziału, w sposób zautomatyzowany. W pierwszym przypadku autorami metadanych są profesjonalni twórcy metadanych (np. bibliotekarze, wydawcy) lub dostawcy treści (np. twórcy stron Web, autorzy publikacji, umieszczający własne treści w repozytoriach). Podstawą oceny jakości metadanych, tworzonych przez ludzi, jest zwykle poziom ich semantycznej i syntaktycznej zgodności z przyjętym schematem metadanych. Do niedawna był to jedyny sposób tworzenia metadanych. Obec-

[1] O rozwoju Internetu może świadczyć wzrost liczby hostów: w 1971 r. połączono ze sobą pierwsze 23 komputery; w 1987 r. było 10 tys. hostów, w 1989 r. – 100 tys.; milion hostów przekroczony został w 1992 r. Obecnie zarejestrowanych jest 760 mln hostów (10 mln w Polsce), z których korzysta 1,8 mld użytkowników.

nie wciąż dominuje on w bibliotekach (w tym cyfrowych), muzeach (również wirtualnych), archiwach i tym podobnych ośrodkach informacji, udostępniających dane o określonej jakości oraz metadane ich dotyczące. Pomimo początkowego braku zainteresowania, coraz częstsze jest także tworzenie przez ludzi metadanych dla dokumentów udostępnianych w Webie, na co wskazuje wzrost stosowania etykiet „słowa kluczowe" i „opis" (*description*) w etykietach <meta> XHTML (Çelik, Meyer, Mullenweg 2005, s. 994). Innym, stosunkowo nowym przykładem jest wykorzystanie w tym celu mikroformatów, takich jak DCMF dla metadanych Dublin Core[2]. Tego typu inicjatywy nigdy nie miały na celu tworzenia metadanych dla wszystkich zasobów Webu – wręcz przeciwnie, dotyczą tylko bardzo niewielkich kolekcji (w stosunku do rozmiarów Internetu, bo istnieją biblioteki cyfrowe udostępniające miliony obiektów, opisanych metadanymi). Dalsze tworzenie metadanych w ten sposób może w przyszłości być utrudnione także z powodu coraz częstszego uwzględniania w bibliotekach zasobów powstających poza tradycyjnymi, scentralizowanymi metodami publikowania, takich jak blogi, osobiste zasoby multimediów i środowisko tagów powstających we współpracy użytkowników Webu. Wszystkie te czynniki powodują, że upowszechnia się tworzenie metadanych oparte na procedurach zautomatyzowanych, których szybkość działania dorównuje szybkości przyrostu treści w Sieci.

Maszynowe przetwarzanie zasobów sieciowych

Automatyczne tworzenie metadanych polega na maszynowym przetwarzaniu zasobów danych. Bibliotekarze i specjaliści od informacji naukowej najczęściej mają do czynienia z automatycznym indeksowaniem, którego głównym celem jest określenie zakresu indeksowanego źródła. Tu warto zwrócić uwagę na różnicę w indeksowaniu treści tzw. semantycznym, realizowanym przez człowieka i program komputerowy. Człowiek indeksujący treść posługuje się zestawem prototypowych pojęć. Przetwarzanie maszynowe opiera się natomiast na identyfikacji zestawu cech. Cechy, wykorzystywane w automatycznym tworzeniu metadanych, stanowią podstawę do podejmowania ocen i decyzji podobnych do podejmowanych przez ludzi, więc powstają z intencją emulacji zachowań podobnych do działań użytkownika-człowieka,

[2] Por. http://www.webposible.com/microformatos-dublincore/microformats_dublin-core.html [dostęp: 2010-10-15]

podczas określania zestawu pojęć. Działania komputerów są jednak inne niż ludzi, gdyż opierają się raczej na zestawie heurystyk lub miarach statystycznych niż właściwym rozumieniu sensu informacji zawartej w obiekcie cyfrowym. W związku z tym konieczne jest zwracanie szczególnej uwagi na kompatybilność obu sposobów oceny treści.

Wyszukiwarki komercyjne stosują automatyczne tworzenie metadanych w dwóch przypadkach. Po pierwsze metadane są tworzone automatycznie, zanim jeszcze rozpocznie się jakiekolwiek wyszukiwanie użytkownika, przy pomocy oprogramowania – pająków, które bez przerwy przeglądają zawartość Webu pobierając i przechowując metadane jej dotyczące w bazach danych wyszukiwarek. Zapytanie użytkownika najpierw jest porównywane z tym apriorycznie przygotowanym zasobem metadanych. Po drugie, w przypadku gdy nie udało się stworzyć odpowiedzi na zapytanie użytkownika na podstawie zgromadzonych zawczasu metadanych, automatycznie i dynamicznie, podczas prowadzenia wyszukiwania przez użytkownika, tworzone są nowe metadane, w wyniku działania algorytmów wyszukiwawczych na otwartym, globalnym zasobie informacyjnym Webu, czyli poza bazą danych wyszukiwarki. Reprezentacje dokumentów w obu sytuacjach na ogół składają się z kilku pierwszych wierszy odszukanych dokumentów (zasobu Web), informacji o lokalizacji (głównie URL) oraz metadanych z etykiety „tytuł".

Jeżeli obiekty cyfrowe, dla których tworzone są metadane, zdefiniujemy bardzo szeroko, jako dowolną jednostkę, formę lub działanie, dla których można zapisać dane kontekstualne (Greenberg 2003, s. 245), to wówczas stwierdzimy, że operacje automatycznego tworzenia metadanych wykonywane są masowo każdego dnia. Przykładem może być automatyczne tworzenie metadanych o wyrażeniach dokumentujących zakupy dokonywane online lub transakcje realizowane za pomocą bankomatów, a także zapisy dotyczące odbytych rozmów telefonicznych. Automatyzacja tych procesów umożliwia przesunięcie ludzi do prac bardziej intelektualnych, również tych związanych z tworzeniem metadanych, a czasem po prostu jest warunkiem ich efektywnego wykonywania.

Automatyczne tworzenie metadanych w najczystszej postaci polega wyłącznie na przetwarzaniu maszynowym. Często są to jednak procesy, w realizację których włącza się ludzi. Mogą oni inicjować odpowiednie procesy, dalej wykonywane automatycznie; wyspecjalizowane oprogramowanie może działać na zasobach metadanych, tworzonych przez człowieka (np. realizując tzw. *mashup* danych bibliograficznych z wielu źródeł, w tym OPAC i innych baz danych); w końcu zdarza się także, że metadane tworzone automatycznie, szczególnie dotyczące za-

kresu dokumentu cyfrowego, są w końcowym etapie ich zestawiania kontrolowane i korygowane przez ludzi. Niektórzy autorzy uważają, że najbardziej efektywnym sposobem tworzenia metadanych jest podejście zintegrowane, łączące udział człowieka i programów komputerowych (Craven 2001). Może to odbywać się na przykład poprzez integrację technik przetwarzania języka naturalnego (Natural Language Processing, NLP) z zastosowaniem słowników kontrolowanych typu tezaurusa (Greenberg, Spurgin, Crystal 2006, s. 5).

Jane Greenberg wskazuje na dwie metody automatycznego tworzenia metadanych dla zasobów cyfrowych: ekstrakcja (*extraction*) metadanych oraz ich zbieranie (*harvesting*) [Greenberg 2004, s. 63]. Ekstrakcja metadanych odbywa się podczas automatycznego wydobywania metadanych z treści dokumentu, wyświetlonego za pomocą przeglądarki internetowej. Treść dokumentu analizowana jest na poziomie podstawowym, co pozwala na tworzenie metadanych ustrukturyzowanych (etykietowanych), składających się na reprezentację obiektu cyfrowego. Można stwierdzić, że ustrukturyzowane metadane są zazwyczaj ekstrahowane z części „body" dokumentu HTML (XHTML). Automatyczna ekstrakcja metadanych może opierać się na skomplikowanych technikach automatycznego indeksowania i algorytmach klasyfikowania treści, które zwiększają jakość metadanych. Część informacji jest łatwa do ekstrakcji dzięki temu, że jej składnia jest dobrze rozpoznawalna. Prostym przykładem może być adres poczty elektronicznej lub strony internetowej URL. Jednak takie przykłady nie są najlepsze, gdyż Internet i jego usługi tworzone były od początku z myślą o wykorzystywaniu komputerów i wymienione elementy zaprojektowane zostały w sposób, ułatwiający ich automatyczne rozpoznawanie i przetwarzanie.

Istnieją także inne elementy danych łatwe do rozpoznania, chociaż w nieco bardziej zawodny sposób. Są to takie wartości, jak kwoty pieniędzy, określenie godziny oraz daty. W tym zakresie kłopoty może sprawiać na przykład istnienie wielu wariantów przedstawienia daty. Część błędów spowodowana jest różnicami kulturowymi, jak na przykład używanie kropki (na kontynencie europejskim) lub przecinka (w krajach anglosaskich) dla oddzielenia części „groszowych" od pełnych jednostek płatniczych.

Często spotykanym, w codziennej praktyce wyszukiwania informacji, przykładem ekstrahowanych metadanych jest informacja o dokumencie Web (będąca w pewnym sensie odpowiednikiem abstraktu), którą wiele wyszukiwarek komercyjnych tworzy dynamicznie podczas

wyświetlania odpowiedzi na zapytanie użytkownika. Różnica pomiędzy tą informacją a „prawdziwym" abstraktem polega na tym, że jest ona tworzona z algorytmicznie pobieranych części dokumentu internetowego (np. kilka pierwszych zdań tekstu lub pierwsze zdanie z każdego akapitu), podczas gdy abstrakt tworzony jest w strukturze intelektualnie logicznej (np. wstęp, metody, wyniki, wnioski). Bez względu na te różnice w obu przypadkach mamy do czynienia z ekstrakcją w takim sensie, że proces realizowany jest na treści dokumentu.

Zbieranie, druga podstawowa metoda automatycznego tworzenia metadanych, ma miejsce, gdy metadane są automatycznie gromadzone z etykiet <meta> znajdujących się w części nagłówkowej (header) kodu źródłowego strony Web w HTML lub też pobierane z zasobów w innych formatach (np. dokumentów MS Word, plików graficznych). Przetwarzanie tego typu opiera się na metadanych, tworzonych zarówno przez ludzi, jak i procesy całkowicie lub częściowo zautomatyzowane (wykonywane przez oprogramowanie). Na przykład oprogramowanie do tworzenia stron Web, takie jak MS Frontpage oraz większość edytorów (w tym MS Word i Adobe) tworzą automatycznie metadane w trakcie powstawania lub modyfikacji dokumentu, zapisując na przykład dane o formacie, dacie utworzenia oraz dacie modyfikacji bez udziału, a nawet wiedzy użytkownika. Podobnie działają urządzenia, takie jak cyfrowe aparaty fotograficzne, które zapisują wspomniane wcześniej metadane w pliku wykonanej fotografii. Tego rodzaju oprogramowanie-generator metadanych może także wspomagać półautomatyczne tworzenie metadanych przez prezentację użytkownikowi (np. autorowi dokumentu lub architektowi stron Web) formularza, przy pomocy którego można, w sposób wspomagany, wprowadzać ręcznie metadane, w rodzaju słów kluczowych lub streszczenia (abstraktu). Oprogramowanie wspomagające automatycznie konwertuje wprowadzone dane do odpowiednich wartości etykiet <meta>, zapisanych w wybranej syntaktyce (np. HTML, XML) i umieszcza je w nagłówku opisywanego dokumentu. Metody te pozwalają na tworzenie metadanych, które nie tylko bezpośrednio ułatwiają wyszukiwanie opisanego obiektu, ale także mogą być zbierane przez generator w celu utworzenia rekordu metadanych ustrukturyzowanych, który z kolei może stać się częścią bazy (meta)danych i służyć jako źródło danych do mashup'u.

Ekstrakcja i zbieranie metadanych są zasadniczymi elementami funkcji realizowanych przez generator metadanych, jednak wciąż brak szczegółowych badań efektywności tych technik dla tworzenia rekordów metadanych. Jeżeli inicjatywy, służące rozwojowi metadanych,

mają odnosić korzyści z możliwości automatycznego przetwarzania danych, niezbędne jest zbadanie wpływu tych metod na jakość metadanych i określenie sposobów ich praktycznego wdrożenia.

Kierunki badań

Badania automatycznego tworzenia metadanych opierają się na pracach naukowych dotyczących automatycznego indeksowania, abstraktowania i klasyfikowania, które rozpoczęły się niedługo po powstaniu pierwszych tekstów elektronicznych, tzn. na początku lat 50. XX wieku. Pierwsze prace w tym zakresie obejmowały głównie tworzenie opisu rzeczowego w postaci deskryptorów/słów kluczowych i abstraktów. Obecnie automatyczne przygotowanie metadanych służy już nie tylko identyfikacji zakresu dokumentu, ale także obejmuje tworzenie wartości dla takich, tradycyjnie uznawanych za formalne, elementów metadanych, jak autor, tytuł, daty (utworzenia, opublikowania, modyfikacji), format i wielu innych. Dodatkowo w Internecie funkcjonują tysiące informacyjnych baz danych, a ich zasoby są często tworzone z użyciem otwartych standardów ułatwiających współdziałanie, takich jak XML. Dzięki temu systemy automatycznego tworzenia metadanych mogą pracować na znacznie większych zasobach, co przyśpiesza przechodzenie z fazy eksperymentów do praktycznych zastosowań.

Badania nad automatycznym tworzeniem metadanych podzielić można na dwa obszary: **badania eksperymentalne**, skupiające się na technikach wyszukiwania informacji i opisu treści zasobów cyfrowych oraz **badania wdrożeniowe** (aplikacyjne), dotyczące głównie rozwoju oprogramowania dla budowy zasobów i narzędzi tworzenia metadanych, stosowanych w działających systemach. Oba obszary omówione są w dalszej części artykułu.

Eksperymenty

Olbrzymie i stale rosnące zasoby cyfrowe dostarczają bogatego materiału do eksperymentów w zakresie badania automatycznego tworzenia metadanych. Naukowcy, pracujący na treściach obiektów cyfrowych w celu tworzenia metadanych, prowadzą badania w dwóch zasadniczych kierunkach: struktury dokumentów i systemów organizacji wiedzy.

W pierwszym przypadku uczeni zidentyfikowali relacje pomiędzy rodzajem, treścią i strukturą dokumentu. Rodzaj dokumentu może na przykład być wnioskowany z gęstości tekstu, która bywa także wyko-

rzystana do przewidywania sposobu działania algorytmu ekstrakcji metadanych dla niektórych rodzajów dokumentów (Greenberg, Spurgin, Crystal 2005, s. 4). W historii badań nad automatycznym indeksowaniem stosowane były dwa podstawowe modele teoretyczne: model wektorowo-przestrzenny i model probabilistyczny. Różnica pomiędzy tymi dwoma modelami jest niewielka, zasadza się na szczegółach ich aplikacji. Metadane mogą być ekstrahowane przy pomocy różnych środków, na przykład z wykorzystaniem maszyn wektorowych dla cech lingwistycznych. Pomyślny przebieg miały eksperymenty badające strukturę dokumentu z wykorzystaniem algorytmów SVM (Support Vector Machine) i DVHMM (Dual Variable Hidden Markov Model) do badania opisów bibliograficznych [Takasu 2003]. Pomyślnie wdrażano także metody heurystyczne.

James Anderson i José Pérez-Carballo przedstawili techniki i strategie automatycznego indeksowania dokumentów tekstowych, w dużej części wypracowane w trakcie serii eksperymentów TREC[3]:

• Podział tekstu na słowa wydaje się tak prostą czynnością, jak zdefiniowanie słowa: jest to ciąg znaków oddzielony spacją lub znakiem przestankowym[4]. Problemy stwarza jednak decyzja o sposobie traktowania znaków przestankowych, takich jak kropki, przecinki, apostrofy lub nawiasy, znajdujących się w obrębie takich jednostek, jak np. symbole chemiczne lub równania matematyczne, które odgrywają zasadniczą rolę w pracach naukowych wielu dyscyplin. Dużym utrudnieniem jest uwzględnianie znaków diakrytycznych, stosowanych w większości języków (poza angielskim). Należy także podjąć decyzję o sposobie traktowania tzw. form złożonych, typu „bardziej pociągający" – czy traktować je jako dwa słowa, czy jedno wyrażenie, a jeśli tak, to w jaki sposób je automatycznie wyróżniać? Dla uniknięcia tych problemów niektórzy badacze odchodzą od wyszczególniania słów na rzecz tworzenia sekwencji znaków – na przykład wszystkich ciągów trój-, cztero- lub pięcioznakowych. Problem powstaje, gdy chcemy uwzględniać cyfry, bo istnieje nieskończona ilość niepowtarzalnych liczb. Inny problem stwarza odróżnianie dużych i małych liter. Podczas obliczania częstotliwości występowania słowa lepiej jest nie odróżniać wielkości liter, natomiast wyrazy pisane dużą literą mogą być przydatne do wyodrębniania nazw własnych. Najprostsze indeksowanie automaty-

[3] TREC – Text REtrieval Conferences (http://trec.nist.gov/) [dostęp: 2010-10-15].
[4] Definicja słowa oparta na spacji i znakach przestankowych jest odpowiednia dla większości systemów alfabetycznych, jednak nie sprawdza się np. w piśmie chińskim.

czne powoduje wyszczególnienie każdego wystąpienia dowolnego słowa; powstałe w ten sposób indeksy mogą być przedstawiane użytkownikowi jako indeks typu KWIC lub KWOC[5]. Oznacza to indeksowanie pełnotekstowe, znane z edytorów tekstów. W zastosowaniach baz danych dodawana jest możliwość tworzenia słów odrzuconych, tzw. stop-listy, dzięki czemu ogranicza się rozmiary indeksu.

• Obliczanie częstości słów i ich wagi. Szybko okazało się, że samo wystąpienie słowa nie świadczy jeszcze o treści lub przeznaczeniu dokumentu. Programy zaczęły liczyć wystąpienia słów w tekście dla określenia częstości ich występowania, co ma lepiej wskazywać na ważne elementy tekstu. Dla grupowania słów różniących się tylko odmianą stosowana jest także analiza morfologiczna, np. przez wskazanie postaci hasłowej słowa i połączenie jej z wszystkimi jego formami, wynikającymi z odmiany. Kolejnym krokiem jest ważenie słów o określonej częstotliwości przez porównywanie ich częstości w danym tekście z częstością występowania w całym zbiorze (np. tekstach danego języka naturalnego lub specjalistycznego). W ten sposób można znaleźć słowa, które w danym tekście (dokumencie) występują z inną częstością niż zwykle. Taka częstość relatywna zwiększa efektywność wyszukiwania: im rzadziej słowo występuje w całym zbiorze dokumentów, tym wyższą wagę otrzymuje jego wystąpienie w konkretnym dokumencie. W ten sposób można obliczyć wagę każdego słowa w dokumencie, a na tej podstawie wagę dokumentu z punktu widzenia zapytania użytkownika.

• W wielu przypadkach wyróżnianie pojedynczych słów nie wystarcza do opisania treści dokumentu. Często połączenia słów oznaczają coś więcej, albo nawet coś zupełnie innego, niż pojedyncze słowa, dlatego bardzo przydatne (choć kosztowne) jest określenie metod i algorytmów identyfikacji fraz w tekście. Polegają one na analizie struktury gramatycznej tekstu w celu identyfikacji części mowy i struktur syntaktycznych.

• Indeksowanie jest zawsze oparte na grupowaniu elementów na podstawie podobieństwa wybranych cech charakterystycznych. Grupowanie (*clustering*) oznacza więc tworzenie klas elementów i/lub przydzielanie elementów do klas. Termin ten stosowany jest dla procesów wykonywanych automatycznie; w przypadku wykonywanych przez człowieka używa się terminu klasyfikowanie. Grupy mogą być tworzone według różnych kryteriów – współwystępowania terminów w dokumentach, au-

[5] KWIC – Keyword in context, KWOC – Keyword out of context.

torstwa, tytułu czasopisma, cytowań. W ten sposób można na przykład oferować wyszukiwanie dokumentów „podobnych" do wskazanego. Techniki automatycznego grupowania służą do obliczania stopnia podobieństwa pomiędzy terminami lub dokumentami. Grupowanie dokumentów jest stosowane do organizowania plików obiektów cyfrowych (grupowanie statyczne) lub w tzw. locie, w celu prezentacji zbioru wyszukanych dokumentów użytkownikowi (grupowanie dynamiczne).

• Interesującym źródłem informacji o wzajemnych relacjach pomiędzy dokumentami są cytowania. Podążanie za cytowaniami, zawartymi w publikacjach uznanych za interesujące dla danego zagadnienia, powoduje tworzenie grupy połączonej cytowaniami. Zastosowanie komputerów znacznie ułatwiło korzystanie z cytowań nie tylko chronologicznie wstecz (kogo cytuje autor znanego nam dzieła), ale także do przodu (kto cytuje autora). Oprócz bardzo przydatnych możliwości wyszukiwawczych, indeksy cytowań wskazują na powiązania pomiędzy dokumentami podobnymi ze względu na wspólny temat, cel, znaczenie. Na podstawie cytowań bibliograficznych także tworzone są grupy – zgodnie z założeniem, że dokumenty posiadające te same opisy w bibliografiach załącznikowych są do siebie podobne. W ten sposób tworzone grupy są statyczne (bibliografie załącznikowe w opublikowanych dokumentach nie zmieniają się). Odwrotna sytuacja ma miejsce w przypadku współcytowania – tu grupa powstaje z dokumentów wspólnie cytowanych w kolejnych, nowych publikacjach. Takie grupy są dynamiczne, gdyż nowych publikacji (z nowymi cytowaniami) wciąż przybywa. Odrębnym zagadnieniem jest problem linków w Webie, które także mogą być uważane za swego rodzaju cytowania (Anderson, Pérez-Carballo 2001b, s. 256-270).

Dla wielu rodzajów dokumentów można także przewidywać ich strukturę, co jest podstawą dla algorytmizowanej ekstrakcji ustrukturyzowanych metadanych. Na przykład artykuły naukowe publikowane w czasopismach naukowych zawierają zazwyczaj standardowe dane, takie jak „tytuł", „autor" oraz „afiliacja autora". Prowadzone były badania służące ekstrakcji tytułu jedynie na podstawie informacji o formacie tekstu, takich jak rozmiar czcionki i umiejscowienie akapitu. Takie podejście znajduje zastosowanie dla dokumentów w językach innych niż angielski (Tonkin, Muller 2008, s. 30). Działania związane określeniem rodzaju jednostek służących do określenia czasu, daty, kwot pieniędzy i nazw własnych w nieustrukturyzowanym tekście nazywane są ekstrakcją jednostek nadrzędnych (ang. *generic entity extraction*). Wiele serwisów stosuje heurystyki, służące do wykrywania tych jednostek.

Metody te można podzielić na:

• wykorzystujące preprogramowane heurystyki; po wstępnym oprogramowaniu heurystyki, do doskonalenia powstałych schematów, wykorzystywana jest ludzka inteligencja, co pozwala na uwzględnienie wyjątków, występujących w językach naturalnych; wiąże się to z koniecznością stałej modyfikacji posiadanego zestawu schematów, powodowanej odkrywaniem i potrzebą uwzględnienia nowych problemów, które w skrajnym przypadku mogą doprowadzić do tego, że system wymknie się spod kontroli;

• służące gromadzeniu konwencji tekstu na podstawie ręcznie etykietowanych danych ćwiczebnych; systemy te zawierają zdefiniowaną strukturę, która może być adaptowana do bieżąco napotykanych wzorców tekstowych. Adaptacja ta uwzględnia parametry oceny zgodne z etykietowanymi dokumentami ćwiczebnymi; w ten sposób można łatwo uwzględniać nowo odkryte warianty przez dodanie do bazy ćwiczebnej kolejnych, ręcznie etykietowanych przykładów;

• heurystyki potrafiące samodzielnie podejmować działania adaptacyjne, wykorzystując dane nieetykietowane; raz wprowadzone techniki autoadaptacyjne mogą funkcjonować autonomicznie dla dużej liczby nieetykietowanych dokumentów. Takie podejście daje dobre efekty przy minimalnym wysiłku manualnym.

Opisane techniki oparte na badaniu (tekstowej) treści dokumentu są znacznie mniej efektywne w zastosowaniu dla zasobów multimedialnych, takich jak wideo, nagrania muzyczne, obiekty graficzne i zestawy danych nietekstowych (np. obliczeń). Poprawna analiza treści tego rodzaju danych jest wciąż przedmiotem aktywnych badań, a jej metody w małym stopniu korzystają z metadanych dotyczących treści. Poszukiwane są inne metody, czego częściowym efektem jest ostatni wzrost zastosowań folksonomii, czyli etykietowania obiektów cyfrowych w oparciu o aktywność społeczności użytkowników. Niestety, swobodne etykietowanie przez ludzi jest użyteczne tylko w przypadku, gdy liczba użytkowników znacznie przekracza liczbę zasobów do etykietowania oraz gdy nie jest wymagane użycie słowników kontrolowanych (patrz dalej) ani standardowych formatów metadanych. W innych sytuacjach Marko Rodriguez, Johan Bollen i Herbert Van De Sompel proponują ekstrapolację (Rodriguez, Bollen, Sompel 2009, s. 7:3) metadanych na podstawie podobieństwa opisywanych dokumentów (chodzi o podobieństwo w zakresie takich cech, jak autorstwo, data publikacji, cytowania). Istnieje duże prawdopodobieństwo, że dokumenty podobne mogą być opisane przy pomocy tych samych, wspólnych metadanych, więc wystarczy wy-

korzystać istniejące metadane dla określonych zasobów, aby opisać nimi podobne zasoby, ale nie opatrzone metadanymi. Można na przykład opisać metadanymi użytkownika, przygotowanymi dla fotografii cyfrowej, wszystkie inne fotografie tego użytkownika wykonane w podobnym czasie (np. w kilkuminutowych odstępach) w tym samym miejscu.

Technologie cyfrowe poważnie zwiększyły dostępność i użyteczność takich systemów organizacji wiedzy, jak ontologie, tezaurusy, systemy klasyfikacyjne, autorytatywne zbiory nazw. W większości były one znane i stosowane wcześniej, jednak w zastosowaniach sieciowych znalazły nowe miejsce, między innymi w automatycznym rozpoznawaniu i wyodrębnianiu metadanych. Rozwój tych narzędzi oraz globalny zasięg Webu spowodowały konieczność budowy rejestrów metadanych specjalnie służących rozpowszechnianiu systemów organizacji wiedzy, takich jak Knowledge System Laboratory (KSL) Ontology Server w Stanford University (http://ksl.stanford.edu/, dostęp: 2010-10-15) oraz rejestr schematów metadanych SCHEMAS (http://www.schemas-forum.org/registry/, dostęp: 2010-10-15) i rejestr elementów Dublin Core (http://dcmi.kc.tsukuba.ac.jp/dcregistry/ dostęp: 2010-10-15) Tego rodzaju zasoby dostarczają kolejnych źródeł do badań automatycznego tworzenia metadanych.

Systemy organizacji wiedzy szczególnie stosowane są podczas tworzenia tzw. metadanych semantycznych, a więc opisujących treść dokumentu [Park, Lu 2009, s. 226]. W tym zakresie wyróżniane są dwa modele. Pierwszy z nich, nazywany etykietowaniem semantycznym z użyciem ontologii, może być stosowany do tworzenia zestawu etykiet semantycznych opisujących treść dokumentu na różnych poziomach strukturalnych. Drugi model, nazywany semantycznym tworzeniem metadanych, ma na celu tworzenie metadanych, które opisują semantycznie treść adnotowanego dokumentu. W tym przypadku można zdefiniować w systemie własną ontologię lub przejąć istniejącą wcześniej (Yang 2009, s. 9710).

Stosowanie ontologii także może rozpoczynać się od opisanej wcześniej ekstrakcji danych z dokumentu. Dane te konwertowane są następnie do metadanych semantycznych w oparciu o posiadaną ontologię. Na podstawie reguł heurystycznych, opartych na ontologii, metadane te są z kolei uzupełniane o elementy ontologii nie występujące wprost w tekście dokumentu.

Ontologie to efektywne narzędzia wspomagające automatyczne tworzenie metadanych, podwyższają one jednak koszty tego przedsięwzięcia, gdyż ich zawartość i struktura wymagają aktualizacji wraz z rozwo-

jem wiedzy. Dlatego też stosowane są metody nie wymagające predefiniowanej ontologii. Wówczas najczęściej ontologia tworzona jest metodą indukcyjną, w procesie generowania metadanych semantycznych, w oparciu o procesy maszynowego uczenia się i zbiór treningowych stron Web, z których pobierany jest wstępny zbiór słów kluczowych.

Istnieją także systemy automatycznej klasyfikacji dokumentów przy pomocy tradycyjnych systemów organizacji wiedzy. Często wykorzystywane są Klasyfikacja Biblioteki Kongresu i Język Haseł Przedmiotowych tej biblioteki. Stosowana tu metoda polega na przetwarzaniu syntaktycznym i wykorzystaniu algorytmów maszynowego uczenia się. Istotną cechą tego rodzaju systemów jest posiadanie przez nie narzędzi oceny metadanych, na podstawie której wspomagane są powtarzalne procesy doskonalenia systemu. W ten sposób system może być dostosowywany do potrzeb otoczenia i udoskonalany w oparciu o wyniki oceny jakości tworzonych metadanych. Wykorzystywane są także inne klasyfikacje, na przykład Ei Thesaurus and Classification Scheme, składający się z dwóch części: tezaurusa terminów technicznych i schematu klasyfikacji (Golub, Lykke 2009, s. 903). Badania wykazały 62-procentową zbieżność indeksowania automatycznego z wykonywanym przez człowieka. W wyniku tych prac zaproponowano także modyfikacje samej klasyfikacji.

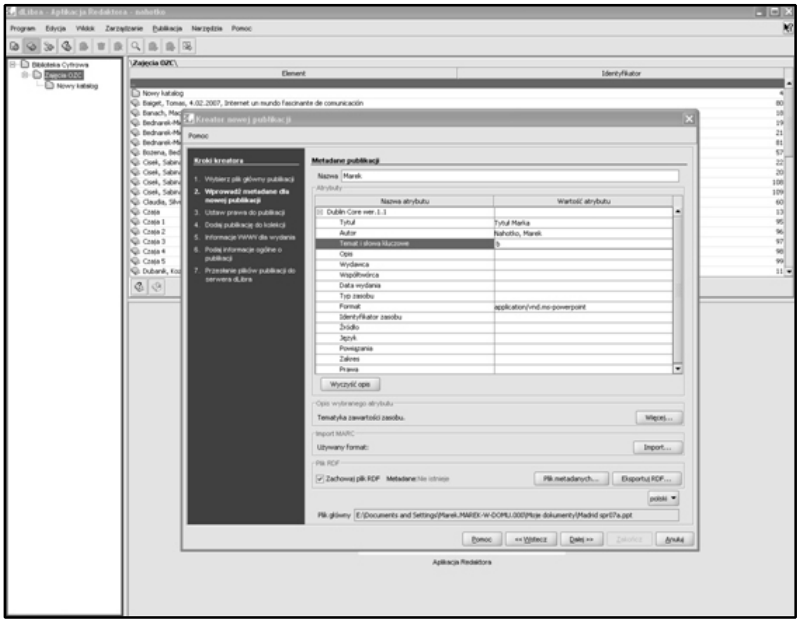
Istnieją również badania służące automatycznej kontroli autorytatywności nazw osobowych przy użyciu KHW Biblioteki Kongresu (Patton i in. 2004). Nazwy w indeksowanym dokumencie ekstrahowane były przez porównanie z kartoteką wzorcową.

Badania eksperymentalne, zajmujące się treścią dokumentu, spowodowały rozwój wiedzy o możliwościach automatycznego tworzenia metadanych. Pewne ograniczenia wynikają z faktu, że badania dotyczą zwykle określonego zakresu, rodzaju i/lub formatu dokumentów i elementów metadanych. Badacze zdają sobie jednak sprawę z tych wad, wynikających z budowy algorytmów dla ograniczonych zastosowań i próbują znaleźć prototypowe narzędzia, pozwalające na tworzenie metadanych z zastosowaniem różnych ontologii (Hatala, Forth 2003). Decyzja o wyborze najlepszej metody wymaga dalszych badań.

Aplikacje

Innym kierunkiem badań jest tworzenie aplikacji wspomagających kreowanie zarówno treści, jak ich metadanych. Do przygotowania metadanych dla zasobów cyfrowych można używać zarówno opro-

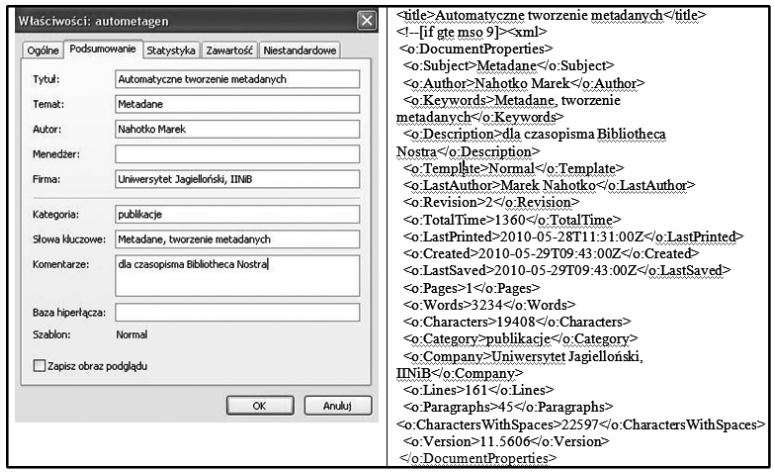
gramowania ogólnego stosowania służącego tworzeniu treści, jak również wyspecjalizowanych narzędzi, nazywanych generatorami metadanych. Aplikacje te stosowane są zazwyczaj przez autorów zasobów lub inne osoby nie posiadające zawodowego wyszkolenia w tworzeniu metadanych. Również zintegrowane systemy biblioteczne posiadają edytory wspomagające tworzenie metadanych, na ogół wyposażone w funkcje pozwalające na kontrolę i optymalizację jakości powstających metadanych. W podobne moduły wyposażone są programy do tworzenia i obsługi bibliotek cyfrowych, takie jak polska dLibra (zob. rys. 1). Tego rodzaju narzędzia przeznaczone są dla profesjonalistów – katalogerów, którzy posiadają odpowiednie przygotowanie i na bieżąco podnoszą kwalifikacje. Możliwości edycyjne zintegrowanych systemów bibliotecznych stanowią jednak odrębne zagadnienie, którego nie będę dalej rozwijał.



Rys. 1. Edytor metadanych systemu dLibra.

Oprogramowanie do tworzenia treści stanowi bardzo zróżnicowaną grupę narzędzi, które łączy jedna cecha – programy te służą do tworzenia dokumentów elektronicznych. Są to wszelkiego rodzaju edytory tekstów z MS Word na czele, ale także Adobe

Acrobat, Macromedia Dreamweaver czy Winamp – czyli każde oprogramowanie, wykorzystywane do tworzenia zasobów, zarówno tekstowych, jak i multimedialnych. W warunkach Webu tego rodzaju oprogramowanie używane jest do tworzenia dokumentu elektronicznego z przeznaczeniem do udostępniania poprzez standardową przeglądarkę Web i związane z nią oprogramowanie. Cyfrowy surogat dokumentu, zawierający dane bibliograficzne (metadane), także może być traktowany jako rodzaj (meta)treści, więc narzędzia typu EndNote Thompsona (http://www.endnote.com/), służące do tworzenia rekordów metadanych (bibliografii załącznikowych), mogą również być uważane za pewną formę oprogramowania tego rodzaju, edytor do tworzenia treści.



Rys. 2. Wypełniony formularz metadanych i metadane osadzone w pliku HTML

Oprogramowanie takie w coraz większym zakresie obsługuje także tworzenie metadanych w różnym trybie: przy pomocy funkcji zautomatyzowanych, częściowo zautomatyzowanych lub wykonywanych przez człowieka. W sposób zautomatyzowany tworzone są takie elementy metadanych technicznych, jak data utworzenia, data modyfikacji, rozmiar, rozdzielczość i format. Często również automatycznie jako twórca dokumentu wskazywany jest właściciel oprogramowania. Niektóre programy pobierają metadane z treści dokumentu, próbując tworzyć w ten sposób metadane opisowe, na przykład Word automatycznie określa tytuł dokumentu w oparciu o zawartość pierwszego wiersza dokumentu. Inne programy pozwalają wyświetlać formularze, wypełniane przez

użytkownika metadanymi. Metadane te mogą być później automatycznie ekstrahowane przez różne aplikacje i konwertowane do wybranego języka kodowania, na przykład XML, co z kolei pozwala na wklejenie ich do opisywanego dokumentu lub umieszczenie w bazie metadanych. Przykład stosowania tej funkcjonalności w Wordzie zawiera rys. 2., gdzie przedstawiony jest formularz (strona lewa), częściowo wypełniany automatycznie, uzupełniany przez użytkownika, zawierający podstawowe metadane o tworzonym dokumencie.

Gdy opisywany dokument zostanie zapisany jako plik HTML, metadane te automatycznie są integrowane z metadanymi z innych formularzy i umieszczane w nagłówku nowego dokumentu (rys. 2., prawa strona). Możliwość zapisania podobnego rodzaju metadanych dają nie tylko komputery, ale też innego rodzaju urządzenia cyfrowe, na przykład aparaty fotograficzne; są takie, które zapisują nie tylko czas wykonania fotografii ale także współrzędne geograficzne fotografowanego miejsca. Służą do tego specjalne formaty metadanych, takie jak standard IPTC (http://ww.iptc.org/).

W coraz większym stopniu do tworzenia metadanych dla zasobów Web stosowane są generatory metadanych. Różnią się one od programów poprzednio opisanych tym, że są one specjalnie i wyłącznie przeznaczone do tworzenia rekordów metadanych. Można je podzielić, ze względu na stosunek pracy wykonywanej przez człowieka i automatycznie, na generatory, które nawet w całości same pobierają metadane z obiektu cyfrowego w sposób zautomatyzowany oraz edytory, które łączą działania automatyczne z przetwarzaniem przez człowieka.

Takie generatory funkcjonują zarówno dla standardowych etykiet <meta>, stosowanych w HTML i XMTML, jak również dla wybranych standardów metadanych ustrukturyzowanych, takich jak Dublin Core. Dla tego schematu dostępnych jest kilka narzędzi, pozwalających zakodować elementy metadanych w wybranej syntaktyce (HTML, XML, RDF), konwertować wyniki do innych schematów, a także kodować wyrażenia tzw. mikroformatów. Często generator automatycznie pobiera metadane ze wskazanej strony Web, przedstawiając rezultaty użytkownikowi w celu dokonania ewentualnych modyfikacji poprzez uzupełnienie wyświetlonego formularza.

Rosnąca liczba tego typu aplikacji jest pozytywnym zjawiskiem, gdyż daje wzrost możliwości efektywnego tworzenia metadanych często przez osoby nie znające tego zagadnienia. Według Jane Greenberg aplikacje te posiadają jednak pewne ograniczenia:

• rzadko pozwalają na stosowanie standardowych funkcji kontroli bibliograficznej, w tym głównie kontroli autorytatywnej i kwalifikacji elementów metadanych;

• rzadko stosowane są w nich rozwinięte techniki i algorytmy automatycznego indeksowania, pomimo, że odpowiednie algorytmy już istnieją;

• tworzone są one w izolacji, co powoduje, że nie uwzględniają poprzednich, pozytywnie zweryfikowanych rozwiązań – jest to częściowo spowodowane brakiem standardów i rekomendowanych funkcji, które mogłyby zostać wykorzystane podczas projektowania aplikacji;

• do badania ich użyteczności i efektywności nie przywiązywano dotąd dostatecznej wagi (Greenberg, Spurgin, Crystal 2005, s.8).

Zakończenie

Oprócz, wcześniej opisanych, problemów technicznych istnieją także sprawy organizacyjne, mające wpływ na automatyczne tworzenie metadanych. Jedną z podstawowych jest posiadanie opracowywanego tekstu (obiektu) w wersji elektronicznej. Tak naprawdę przydatność procedur automatycznych widoczna jest w ich zastosowaniu do masowych zasobów sieciowych – ze wskazaniem na zasoby internetowe, których inaczej nie sposób zindeksować.

Istnienie wersji elektronicznej łączy się z problemem doboru formatów. Można ogólnie powiedzieć, że dokumenty w niektórych formatach znacznie łatwiej poddawać komputerowej obróbce, niż w innych, gdyż format wpływa na poziom dostępności dokumentu. Należy więc uwzględnić te różnice już na etapie tworzenia dokumentów.

Istnieją także zagadnienia prawne, dotyczące wykorzystania lub analizy pełnych tekstów dokumentów dla ekstrahowania z nich surogatów. Po pierwsze mogą wystąpić problemy z akceptacją żądania dostępu do dokumentu przez posiadający go serwis. Odrębnym zagadnieniem jest publikacja rekordu metadanych. W pewnych przypadkach wskazane jest bowiem zachowanie pełnego tekstu dokumentu przez system indeksujący (jak to robi Google) w celu umożliwienia jego kolejnego przetwarzania i częściowej analizy treści.

Trudności sprawiać może ocena jakości rekordu metadanych utworzonego automatycznie. Stosując metody komputerowe, szczególnie oparte na statystyce, można oszacować stopień poprawności osiągniętych wyników. Ujawnianie tych informacji użytkownikom systemu nie jest jednak powszechnie przyjętą praktyką. Z punktu widzenia

użytkownika metadanych informacja ta jest więc tracona i nie jest wykorzystywana podczas stosowania metadanych.

Również język dokumentu i jego lokalizacja mogą powodować kłopoty podczas automatycznego tworzenia metadanych. Wiele z wymienionych narzędzi i metod (algorytmów) jest przygotowana do pracy w określonym języku, zazwyczaj angielskim.

Jak piszą James Anderson i José Pérez-Carballo, automatyczne indeksowanie dokumentów funkcjonuje poprawnie, można nawet powiedzieć, że daje równie dobre rezultaty, jak indeksowanie realizowane przez człowieka, jest tylko inne (Anderson, Pérez-Carballo 2001a, s. 236). Jest ono jednocześnie znacznie szybsze i tańsze niż tworzenie metadanych przez ludzi na podstawie analizy intelektualnej i na tyle wydajne, że może być wykorzystane do indeksowania nawet tak olbrzymich, a jednocześnie tak dynamicznych zasobów, jak te funkcjonujące we współczesnym Webie. Według opinii Bożenny Bojar, poza stosowaniem różnego rodzaju metod przetwarzania języka naturalnego, nie ma innej alternatywy dla umożliwienia efektywnego wyszukiwania we współczesnych zasobach informacyjnych, tak wielkich, że z chaosu tych zasobów, w wyniku samoorganizacji, wyłania się nowy porządek (Bojar 2009, s. 19-23), który nie może być już opisany przy pomocy tradycyjnych narzędzi, służących do tworzenia metadanych przez ludzi.

Bibliografia

Anderson J., Pérez-Carballo J. (2001a), *The nature of indexing : how humans and machines analyze messages and texts for retrieval.* Part I: *research, and the nature of human indexing.* „Information Processing and Management" vol. 37, s. 321-245.

Anderson J., Pérez-Carballo J. (2001b), *The nature of indexing : how humans and machines analyze messages and texts for retrieval.* Part II: *machine indexing and the allocation of human versus machine effort.* „Information Processing and Management" vol. 37, s. 255-277.

Bojar B. (2009). *Języki informacyjno-wyszukiwawcze wczoraj, dziś… czy jutro?* „Zagadnienia Informacji Naukowej" nr 1, s. 3-24.

Çelik T., Meyer E., Mullenweg M. (2005). *XHTML Meta Data Profiles.* W: Proc. of 14th International Conference of the World Wide Web Consortium (WWW2005), Chiba, Japan, 10-14 May 2005. Pod red. A. Ellis, T. Hagino. New York, s. 994--995.

Craven T. (2001), *Description meta tags in public home and linked pages* [online]

„Libres" vol. 11 nr 2. [dostęp: 2010-05-15]. Dostępny w World Wide Web: http://libres.curtin.edu.au/LIBRE11N2/craven.htm

Golub K., Lykke M. (2009). *Automated classification of Web pages in hierarchical browsing.* „Journal of Documentation" vol. 65 nr 6, s. 901-925.

Greenberg J. (2004). *Metadata extraction and harvesting: a comparison of two automatic metadata generation applications.* „Journal of Internet Cataloging" vol. 6 nr 4, s. 59-82.

Greenberg J. (2003), *Metadata and the Word Wide Web.* W: The encyclopedia of library and information science. 2nd ed. Pod red. M. Drake, vol. 72. New York, s. 244-261.

Greenberg J., Spurgin K., Crystal A. (2006), *Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions.* „Intern. Journal on Metadata, Semantics and Ontologies" vol. 1 nr 1, s. 3-20.

Greenberg J., Spurgin K., Crystal A. (2005), *Final report for the AMeGA (Automatic Metadata Generation Applications) Project* [online]. [dostęp: 2010-05-12]. Dostępny w World Wide Web: http://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf

Hatala M., Forth S. (2003), *A comprehensive system for computer-aided metadata creation.* W: Proc. of 12th International Conference of the World Wide Web Consortium (WWW2003), Budapest, 20-24 May 2003. Pod red. G. Hencsey, B. White. New York, s. 334.

Nahotko M. (2010), *Komunikacja naukowa w środowisku cyfrowym.* Warszawa.

Park J., Lu C. (2009), *Application of semi-automatic metadata generation in libraries: types, tools and techniques.* „Library & Information Science Research" vol. 31, s. 225-321.

Patton M. [i in.] (2004), *Toward a metadata generation framework: a case study at the John Hopkins University.* „D-Lib Magazine" [online]. 2004, Vol. 10 nr 11 [dostęp: 2010-05-29]. Dostępny w World Wide Web: http://www.dlib.org/dlib/november04/choudhury/11choudhury.html.

Rodriguez M., Bollen J., Sompel H. (2009), *Automatic metadata generation using associative networks.* „ACM Transactions on Information Systems" vol. 27 nr 2, art. 7.

Takasu A. (2003), *Bibliographic attribute extraction from erroneous references based on a model.* W: Proc. of the 3rd ACM/IEEE conference on digital libraries JCDL. Pod red. L. Delcambre, G. Henry, C. Marshall. Washington, s. 49-60.

Tonkin E., Muller H. (2008), *Keyword and metadata extraction from pre-prints.* W: Open scholarship authority, community and sustainability in the age of Web 2.0. Proc. of the 12th Intern. Conference on Electronic Publishing ELPUB 2008, Toronto 25-27 June 2008. Pod red. L. Chan i S. Mornati. Torino, s. 30-44.

Yang H. (2009), *Automatic generation of semantically enriched web pages by a text mining approach.* „Expert Systems with Applications" vol. 36, s. 9709-9718.

M. Nahotko ***Automatic metadata creation***
Summary

Because of the enormous Web resources and their rapid increase it has become impossible to develop and indexing them by traditional methods - by trained cataloguers. Therefore are more and more widely methods of automatic metadata creation used, both on the formal characteristics (author, title, source) and the content of text documents. The article describes the latest developments in this field, both experimental studies and implementations.