Zeszyty Naukowe
Wyższej Szkoły Bankowej w Poznaniu
Nr 32/2011

Marek Nahotko

Uniwersytet Jagielloński
Instytut Informacji Naukowej i Bibliotekoznawstwa

Publikowanie semantyczne

Streszczenie. *Dzięki rozwojowi technologii Webu możliwa jest budowa semantycznego środowiska publikowania czasopism naukowych i poszczególnych artykułów. Szczególną wartość ma uzupełnianie tekstu artykułu o dane z badań i metadane oraz organizowanie interaktywnego dostępu do treści. Tego typu działania, będące elementami ideologii Semantycznego Webu, już obecnie są realizowane przez wydawców czasopism z nauk ścisłych, techniki i medycyny (STM). Istotną rolę w tych pracach może mieć także automatyczne przetwarzanie tekstu; dzięki niemu działania te staną się możliwe, a nawet rutynowe. Najważniejsze role w tym procesie mają do odegrania wydawca, redaktor oraz autor; role te jednak ulegają poważnym modyfikacjom w stosunku do publikowania tradycyjnego.*

1. Wprowadzenie

Efektywna komunikacja wyników badań naukowych ma zasadnicze znaczenie dla realizacji cyklu badań naukowych, składającego się z takich etapów, jak: formułowania hipotezy, realizacji badań, interpretacji wyników oraz ich publikacji, przy czym wyniki jednego projektu badawczego stanowią podstawę dla stawiania hipotez w kolejnym projekcie. Od czasu pojawienia się pierwszych czasopism na-

ukowych artykuł w periodyku wciąż jest głównym sposobem rozpowszechniania i archiwizacji nowych wyników badań. Pomimo zmian technologicznych, jakie czasopisma naukowe przeszły w ciągu ostatnich dziesięciu lat, zasadnicza struktura artykułu naukowego pozostaje niezmieniona, co powoduje, że możliwości oferowane przez środowisko Web nie są w pełni wykorzystywane.

W ostatnich latach rozwój technologii, stosowanych w Web stał się przyczyną rewolucyjnych zmian, zachodzących w środowisku naukowym. Nowe sposoby komunikacji cyfrowej pomiędzy uczonymi (archiwa reprintów, blogi, wideokonferencje itp.) powodują zmiany ról głównych aktorów (uczonych, wydawców, bibliotekarzy, grantodawców) w cyklu publikowania naukowego. Zamierzeniem autora artykułu jest jednak ograniczenie się do przedstawienia zmian zachodzących w artykułach publikowanych w czasopismach naukowych, w szczególności w zakresie rozbudowy ich semantyki. Chodzi głównie o takie kodowanie znaczenia, które pozwala na automatyzację procesów użytkowania publikacji. Tego typu wzbogacanie semantyki już obecnie znalazło praktyczne zastosowanie w części czasopism.

Dokumenty publikowane są w Webie głównie z myślą o ludziach jako odbiorcach ich treści. Twórca idei Webu, Tim Berners-Lee, jest także autorem wizji „Webu danych", którego struktura umożliwi komputerowe wspomaganie ludzi w wykonywaniu zadań związanych z eksploracją informacji, integracją danych cyfrowych i zarządzaniem wiedzą. Idee te, nazwane „Semantycznym Webem" lub Webem 3.0, pomimo początkowych niepowodzeń, zaczynają obecnie przynosić realne korzyści. Jak się okazało, do interpretacji informacji, zawartych w publikacjach, nie potrzeba złożonych procedur sztucznej inteligencji, ale raczej, jak twierdzi Berners-Lee, interpretacja ta opiera się na zdolności komputerów do rozwiązywania dobrze zdefiniowanych problemów, dzięki realizacji dobrze zdefiniowanych operacji na dobrze zdefiniowanych danych; takie podejście może gruntownie zmienić sposób tworzenia i rozpowszechniania wiedzy naukowej[1]. Sudeshna Das określa Semantyczny Web jako Web 2.0, w którym dane są ze sobą powiązane, a ich semantyka znana[2]. W uproszczeniu oznacza to, że w Semantycznym Webie strony Web posiadają znaną strukturę – schemat, podobnie jak ma to miejsce w bazach danych (np. bibliotecznych OPAC[3]). Schemat jest oparty na modelu grafu, może być opublikowany i rozpowszechniony. W efekcie strony Web mogą być serwisami Webu i/lub repozytoriami danych. Semantyka takich danych i serwisów jest dobrze zdefiniowana, a informacja wyrażona w formacie zrozumiałym dla maszyny może być przetwarzana przez agentów programowych. Obecnie zalety te stają się coraz szerzej dostępne, między innymi dzięki tworze-

[1] T. Berners-Lee, J. Hendler, O. Lasilla, *The Semantic Web, „Scientific American" 2001*, vol. 284, nr 5, s. 37.

[2] S. Das et al., *Scientific publications on Web 3.0*, w: *13th International Conference on Electronic Publishing ElPub 2009*, Milan 2009, s. 108.

[3] OPAC – *Online Public-Access Catalog*.

niu czytelnych maszynowo metadanych, stosowaniu uzgodnionych standardów, pozwalających na komputerowe wspomaganie wydobywania i integracji informacji. Do publikowania semantycznego przydatne jest także wykorzystanie powszechnie stosowanych technologii Webu, takich jak hiperlinki i interaktywność dokumentów Web.

David Shotton publikowaniem semantycznym określa wszystkie działania, które zwiększają możliwości zrozumienia treści opublikowanego artykułu naukowego, ułatwiają jego automatyczne wyszukanie, umożliwiają tworzenie powiązań z podobnymi semantycznie artykułami, dają dostęp do danych zawartych w artykule lub ułatwiają integrację danych z wielu artykułów[4]. Oznacza ono między innymi wzbogacanie artykułów metadanymi, umożliwiającymi automatyczne przetwarzanie i analizę, zwiększającymi weryfikowalność opublikowanych informacji oraz umożliwiającymi jej automatyczne gromadzenie i indeksowanie. Tego typu udogodnienia semantyczne podnoszą wartość artykułów w czasopismach dzięki ułatwieniu ekstrakcji zawartej w nich informacji i wiedzy. Możliwe jest także tworzenie nowych usług, w tym służących integracji informacji pochodzącej z tak udoskonalonych artykułów, co daje dodatkowe możliwości biznesowe dla ich wydawców. Równie istotne są korzyści użytkowników, wynikające z szybszego, łatwiejszego i pełniejszego dostępu do informacji.

Dodatkowa wartość tworzona jest dzięki integracji informacji z wielu artykułów, co umożliwia ich meta-analizę i/lub łączenie, mieszanie (*mashup*) z danymi z innych źródeł (np. z map cyfrowych). Publikowanie semantyczne wymaga wydatkowania dodatkowego wysiłku podczas procesu publikowania, jednak korzyścią jest znaczny wzrost efektywności badań naukowych. Niektóre z nich inaczej nie byłyby zapewne w ogóle realizowane (np. związane z modelowaniem matematycznym), gdyż wysoki koszt analiz tradycyjnie opublikowanych danych bardzo ogranicza ich stosowanie.

2. Zalety i wady obecnych czasopism online

Współczesne sposoby publikowania czasopism elektronicznych charakteryzują się wieloma zaletami w stosunku do wcześniejszych metod tworzenia i rozpowszechniania czasopism drukowanych. Niestety, wciąż istnieją także wady warte wyeliminowania. Zalety i wady czasopism elektronicznych przedstawiłem w jednej z moich wcześniejszych publikacji[5]. Tutaj skupię się tylko na niektórych, istotnych dla publikowania semantycznego.

---

[4] D. Shotton, *Semantic publishing: the coming revolution in scientific journal publishing*, „Learned Publishing" 2009, vol. 22, nr 2, s. 87.

[5] M. Nahotko, *Naukowe czasopisma elektroniczne*, Wyd. SBP, Warszawa 2007, s. 18-20.

Do zalet zaliczyć można eliminację procesu przygotowania składu drukarskiego; czasopismo w pełni online nie wymaga także czasochłonnego procesu druku i fizycznej dystrybucji[6], co skraca czas niezbędny dla realizacji cyklu komunikacji naukowej. Przyśpieszeniu ulegają także inne procesy, jak na przykład recenzowanie, dzięki możliwości pominięcia potrzeby wysyłania przesyłek tradycyjną pocztą pomiędzy redakcją a recenzentami. W efekcie publikacje elektroniczne produkowane są szybciej i taniej oraz łatwiej dystrybuowane. Publikowanie elektroniczne stworzyło możliwość stosowania nowych modeli dystrybucji, będących alternatywą dla tradycyjnej prenumeraty pojedynczych tytułów, takich jak Big Deal[7], *pay per view*[8] czy czasopisma wirtualne oraz spowodowało powstanie nowych serwisów oferujących rozszerzone funkcje wyszukiwania, przeglądania i ekstrakcji informacji. Uczonym-autorom e-publikowanie przynosi oczywiste korzyści: umożliwia dostarczanie i modyfikację tekstu w formie elektronicznej online, zmniejsza ograniczenia co do objętości tekstu i pozwala na dołączanie dodatkowych załączników (np. tzw. surowych danych z badań), a więc pozwala na publikowanie materiałów, które wcześniej nie wychodziły poza laboratorium badacza. Treść artykułów może być modyfikowana między innymi dzięki dyskusjom toczonym z czytelnikami w Webie, stającymi się specyficznym rodzajem recenzowania. Uczeni-czytelnicy korzystają z większej dostępności do czasopism online, rewolucjonizujących zachowania czytelnicze, gdyż nie ma już potrzeby dostosowywania własnego rytmu pracy do czasu pracy biblioteki czy innej instytucji. Funkcjonowanie serwisów wyszukiwawczych i powiadamiania (RSS) pozwala pomijać konieczność przeglądania spisów treści, a linki przez DOI wspomagają eksplorację i gromadzenie wiedzy; dostępne są dane wraz z narzędziami do ich analizy i opracowania.

Korzyści z publikowania online wiążą się także koniecznością ponoszenia pewnych kosztów. Wydawnictwa i biblioteki potrzebują personelu o nowych kwalifikacjach związanych ze stosowanymi technologiami, zdolnego do ciągłej pogoni za nowymi osiągnięciami w tym zakresie. Technologie te są kosztowne, więc ich implementacja powoduje komasację rynku wydawniczego; duże wydawnictwa stają się coraz większe, małe znikają. Sposób, w jaki zmiany wpływają na autorów może zaskakiwać; pomimo większego i łatwiejszego dostępu

---

[6] To spowodowało już zanik pewnych tradycyjnych zawodów związanych z publikowaniem, na przykład zecera.

[7] Część nowych modeli jest krytykowana, na przykład Big Deal jest dobrym rozwiązaniem z punktu widzenia wydawców, krytykowany jest jednak przez bibliotekarzy za utrudnianie prowadzenia przez nich świadomej polityki gromadzenia zbiorów. Polega on na oferowaniu bibliotekom całych pakietów czasopism, w których znajdują się tytuły bardziej i mniej pożądane przez użytkowników.

[8] Model *pay per view* polega na pobieraniu płatności za dostęp do poszczególnych artykułów, bez potrzeby prenumeraty czasopisma.

online do czasopism stosują oni tzw. *tunnel vision*, cytując w swoich pracach artykuły mniej zróżnicowane oraz bieżąco publikowane (starsze publikacje są eliminowane) w stosunku do tego, co miało miejsce w okresie publikacji drukowanych. Wynika to z łatwości dotarcia do odpowiednich, bieżących publikacji, wystarczających do gromadzenia niezbędnej informacji, niż w okresie publikacji drukowanych[9]. Nie bez znaczenia jest też zapewne większa dostępność bieżących źródeł oraz wzrastające tempo zmian i powstawania nowych idei w nauce. Czytelników dotyczy inny problem, gdyż mają oni dostęp do znacznie większej liczby czasopism niż kiedykolwiek wcześniej, co powoduje coraz większy zalew danych i informacji (*data deluge*). Wyznacza to również nowe role bibliotekarzom, zmuszonym do ponownego przemyślenia swojego miejsca w świecie informacji cyfrowej.

W zakresie standardów, stosowanych w publikowaniu elektronicznym, wciąż występują pewne niedostatki. Wersje online artykułów są zazwyczaj prezentowane jako standardowe strony HTML, na ogół bez specjalnego kodowania, lub też dostarczane jako pliki PDF. Ten ostatni format, naśladujący stronę drukowaną tekstu, jest zrozumiały i łatwy do stosowania (szczególnie drukowania) przez użytkowników. Jest on jednak zaprzeczeniem idei Webu, zachowując statyczność druku i utrudniając stosowanie zautomatyzowanych narzędzi, wykorzystywanych między innymi do linkowania informacji pomiędzy artykułami.

Coraz więcej artykułów jest opatrywanych identyfikatorami DOI, ale wciąż rzadko są one wykorzystywane do tworzenia linków między artykułem cytującym i cytowanymi w bibliografii załącznikowej. Wszystko, co umożliwia dostęp do przydatnych artykułów przez jedno kliknięcie, bardzo zwiększa użyteczność czasopisma. W związku z tym coraz więcej baz danych artykułów, tworzonych i zarządzanych przez poszczególnych wydawców, przyjmuje formę indeksu cytowań, pozwalając na badanie cytowań w obie strony (od i do czytanego artykułu) przez jedno kliknięcie na identyfikator typu DOI.

Niezaprzeczalną korzyścią, związaną z czasopismami elektronicznymi, jest możliwość dołączania dodatkowych materiałów w formie elektronicznej, np. danych z badań[10]. Problemy sprawia brak wspólnie uzgodnionych i stosowanych standardów, niedostateczna kontrola jakości, niemożność dotarcia do tych materiałów przez wyszukiwarki w celu ich indeksowania. W konsekwencji wykorzystanie cennych materiałów uzupełniających zasadniczą treść artykułu naukowego jest niewystarczające.

---

[9] J. Evans, *Electronic publication and the narrowing of science and scholarship*, „Science" 2008, nr 321, s. 396.

[10] Ch. Borgman, *Data, disciplines and scholarly publishing*, „Learned Publishing" 2008, vol. 21, nr 1, s. 31.

Oprócz kilku wyjątków, czasopisma online nadal nie stosują semantycznego kodowania tekstów (*markup*)[11], które mogłoby ułatwić zrozumienie znaczenia zawartych w nich treści. Poważnym problemem jest brak metadanych, które powinny być „zaszyte" w tekście i dostępne do wyszukiwania. W tym zakresie można jeszcze wiele zmienić: proste metadane ułatwiają odszukanie artykułu, a odpowiednie kodowanie semantyczne wyników badań, opisanych w artykule, może pozwolić na zautomatyzowane metawyszukiwanie; w efekcie zmienia się sposób prowadzenia prac naukowych.

3. Publikowanie semantyczne

W tej części artykułu przedstawione zostaną aktualnie prowadzone prace, służące zwiększeniu siły semantycznej publikacji naukowych oraz możliwości występujące w tym zakresie, które mogą być wykorzystane w przyszłości. Większość opisanych prac ma charakter eksperymentalny. Omówione zostaną także nowe role aktorów uczestniczących w procesie publikowania (ludzi i oprogramowania).

3.1. Obecnie realizowane prace

Wydawcy coraz częściej stosują technologie i zasady pracy, zbliżające ich produkt do publikowania semantycznego. Najbardziej symptomatyczne są następujące prace:

1. Część wydawców, np. PLoS[12] udostępnia już, oprócz tekstów artykułów w PDF, służących do wyświetlania i drukowania, także tekst zakodowany w XML; kod można skopiować na lokalny komputer w celu dalszego wykorzystania, na przykład do łączenia z innymi tekstami.
2. Wydawcy udostępniają nieprzetworzone dane z badań; przykładem, co prawda spoza nauki, może być baza OECD pn. SourceOECD[13], zawierająca dane numeryczne, które mogą być pobierane jako skoroszyty Excela. Podobnie PLoS udostępnia dane w Excelu o poziomie wykorzystania publikowanych artykułów. Możliwe jest też tworzenie graficznych wizualizacji danych, np. wielowymiarowych struktur kryształów w czasopiśmie Acta Crystallographica Section E[14] (obecnie czasopismo Open Access).

[11] Chodzi głównie o kodowanie tekstów w XML i podobnych językach kodowania.

[12] Public Library of Science (PLoS) jest wydawcą czasopism naukowych Open Access o najwyższej jakości naukowej z zakresu biologii i medycyny (http://www.plos.org/).

[13] Dostępne pod adresem: http://www.sourceoecd.org/.

[14] Dostępne pod adresem: http://journals.iucr.org/e/journalhomepage.html.

3. Coraz odważniej stosowane są multimedia (grafika, dźwięk i wideo). Ciekawym przykładem jest New England Journal of Medicine[15] (NEJM), który proponuje użytkownikom m.in. konkurs na właściwą diagnozę medyczną na podstawie wyświetlanych obrazów zmian chorobowych, a także publikacje audio i wideo.
4. Podejmowane są eksperymenty dotyczące strukturyzacji tekstu artykułu, bardziej odpowiedniej dla środowiska cyfrowego. Prace takie, oprócz wspomnianych NEJM i PLoS[16] podjęli także wydawcy komercyjni, np. Elsevier dla czasopisma Cell[17] jako części projektu Articles of the Future[18].
5. Nature Publishing Group dla swoich klientów i uczonych udostępnia bezpłatnie narzędzie do zarządzania zakładkami (ang. *bookmarking*) o nazwie Connotea[19]; umożliwia ono szybkie zapisywanie, organizację, etykietowanie i rozpowszechnianie odsyłaczy bibliograficznych. Podobnymi narzędziami są CiteULike[20], Zotero[21] i Mendeley[22]. Umożliwiają tworzenie społecznościowych zakładek na wzór najbardziej znanego systemu tego typu, jakim jest del.icio.us[23]. Twórcy Zotero i Mendeley mają poza tym znacznie większe ambicje; serwisy te mają całościowo organizować środowisko pracy uczonego, pozwalając na organizację literatury naukowej w formie elektronicznej, w tym automatyczne indeksowanie (tworzenie metadanych) artykułów naukowych, tworzenie notatek powiązanych z gromadzonymi artykułami i wspomaganie tworzenia i „stylizowania" bibliografii. Możliwe jest tworzenie statystyk wykorzystania własnych i cudzych prac.
6. Wiele redakcji dzieli publikowane abstrakty na sekcje, zawierające części dotyczące artykułu, typu: założenia, metodologia, główne tezy i wnioski. Kolejnym krokiem są tzw. strukturalne abstrakty cyfrowe, zawierające czytelne dla maszyny streszczenia najważniejszych danych i wniosków z artykułu, zazwy-

---

[15] Dostępne pod adresem: http://beta.nejm.org/.

[16] D. Shotton et al., *Adventures in semantic publishing: exemplar semantic enhancements of a research article*, „PLoS Computational Biology" 2009, vol. 5, nr 4 (online), dostępne na stronie: http://www.ploscompbiol. org/article/info:doi/10.1371/journal.pcbi.1000361, dostęp: 17 marca 2010 r.

[17] Dostępne pod adresem: http://beta.cell.com/.

[18] Projekt jest kontynuacją wcześniejszych działań Elseviera, takich jak Article 2.0 i Elsevier Grand Challenge.

[19] Dostępne pod adresem: http://www.connotea.org/about.

[20] Dostępne pod adresem: http://www.citeulike.org/.

[21] Dostępne pod adresem: http://www.zotero.org/.

[22] Dostępne pod adresem: http://www.mendeley.com/.

[23] Serwis del.icio.us (http://delicionus.com) pozwala na tworzenie zakładek w wyszukiwarce ułatwiających szybki powrót do strony opisanej w zakładce, a więc jest to mechanizm znany z Ulubionych, jednak różnica polega na tym, że zakładki przechowywane są w Internecie wraz z informacją (metadanymi, np. jednowyrazowymi etykietkami) o sposobach opisu danej strony przez innych internautów i popularności strony.

czaj kodowane w XML. Dla tekstów z zakresu biologii i medycyny powstały w tym celu specjalne standardy, takie jak MIMIx (Minimum Information required for reporting a Molecular Interaction eXperiment), który jest jedną z wielu ontologii zawierających tzw. minima informacji o przeprowadzonym eksperymencie badawczym[24], utworzonych w projekcie MIBBI[25].

7. Niektóre redakcje czasopism naukowych realizują eksperymenty dotyczące semantycznego kodowania tekstu. Na przykład w czasopiśmie Molecular BioSystems[26] istnieje dostęp do rozszerzonej wersji HTML każdego artykułu, która pozwala na podświetlanie w tekście terminów z wielu słowników kontrolowanych i ontologii, takich jak: Compendium of Chemical Terminology, Gene Ontology (GO), Sequence Ontology, Cell Type Ontology i in. Po kliknięciu na podświetlony termin użytkownik otrzymuje dodatkowe informacje i linki (definicje, synonimy, linki do innych artykułów, patentów, wzorów strukturalnych, opisu związków chemicznych przedstawionych w CML[27]).
8. Science Collaboration Framework (SCF) jest oprogramowaniem udostępnianym na podstawie licencji GNU, za pomocą którego wykonano serwis StemBook[28], zawierający zbiór recenzowanych artykułów z zakresu biologii, dostępnych online. Teksty tworzone są w XML, skąd konwertowane są do XHTML i PDF. Podstawą tej modernizacji są tzw. adnotacje semantyczne, nie służące co prawda do rozwiązywania problemów biologicznych, ale pozwalające odnajdywać fakty i źródła. W systemie wykorzystywane są ontologie i słowniki kontrolowane, takie jak GO. System wyszukuje w artykule terminy pobierane z GO i porównuje zawartość tekstu i ontologii; w rezultacie proponowane są terminy do semantycznej adnotacji. Powstał skomplikowany graficzny interfejs użytkownika, przedstawiający wyniki wyszukiwań w formie grafów RDF.
9. Prace prowadzone w ramach PLoS. W eksperymentalnych wydaniach czasopisma PLoS Neglected Tropical Diseases[29] wprowadzono takie elementy semantyki tekstu, jak: żywe DOI i hiperlinki, semantyczne kodowanie terminów z linkami do informacji uzupełniającej, obliczenia interaktywne, bibliografie z możliwością decyzji o uszeregowaniu opisów według wybranych kryteriów. Tworzone jest streszczenie dokumentu, zawierające zestawienia statystyczne, chmury etykiet i hierarchiczne drzewka etykiet kodowanych jednostek

---

[24] Ch. Taylor et al., *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project*, „Nature Biotechnology" 2008, vol. 26, nr 8, s. 891.

[25] Dostępne pod adresem: http://www.mibbi.org/.

[26] Dostępne pod adresem: http://www.rsc.org/Publishing/Journals/MB/.

[27] CML – Chemical Markup Language, odmiana XML dla publikacji chemicznych (http://cml.sourceforge.net/).

[28] Dostępne pod adresem: http://www.stembook.org/.

[29] Dostępne pod adresem: http://www.plosntds.org/home.action.

(*Tag Trees*) oraz analizy cytowań w obrębie artykułu (*Citations in Context*). Dodawany jest ustrukturyzowany abstrakt, zawierający podstawowe dane o chorobach tropikalnych, stanowiących tematykę czasopisma. Dołączane są arkusze kalkulacyjne, zawierające dane z artykułu wraz z informacją o ich pochodzeniu, w połączeniu z danymi z innych artykułów. Dostępne są metadane zarówno o wybranym artykule, jak i o artykułach przez niego cytowanych.

3.2. Zadania poszczególnych aktorów

Warto zastanowić się nad zadaniami, jakie mają do wypełnienia poszczególni aktorzy, uczestniczący w procesie publikowania, na drodze do pełnego zastosowania zasad publikowania semantycznego. Poniżej omówię role podstawowych aktorów: wydawców, uczonych, redaktorów, a także procesów możliwych do zautomatyzowania, realizowanych przez aktorów programowych. Grupy te tworzą społeczności, funkcjonujące w ramach Nauki 2.0, będącej częścią Webu 2.0, a często także Semantycznego Webu, posługując się opisanymi wcześniej narzędziami Webu społecznościowego.

1. Wydawcy powinni uwolnić informacje i dane. Obecnie trwają intensywne prace nad stworzeniem modelu pozwalającego połączyć możliwość uzyskiwania komercyjnych zysków z otwartym dostępem do publikacji i towarzyszących im danych dla użytkowników końcowych[30]. Rozwija się także ruch Open Access wydawnictw niekomercyjnych; swój udział w nim mają również bibliotekarze, tworzący biblioteki cyfrowe. Wydawcy zawsze czerpali pośrednie korzyści z bezpłatnego udostępniania metadanych, np. bibliograficznych, które stanowiły opis końcowego produktu, będącego przedmiotem ich bezpośrednich działań komercyjnych – artykułu lub książki naukowej. Przykładami mogą być bazy danych wydawców bezpłatnie udostępniające opisy wraz z abstraktami, co przynosi konkretne zyski dzięki temu, że użytkownicy mogą łatwo odnaleźć potrzebne im publikacje (a następnie zapłacić za ich udostępnienie). Podobnie Amazon[31] udostępnia pełne metadane publikacji oferowanych w tej e-księgarni, co daje korzyści wynikające z możliwości wejścia do ich serwisu z dowolnego miejsca w sieci. Oprócz dostępu do tekstów naukowych użytkownicy powinni mieć możliwość dotarcia do danych przedstawionych w tabelach i na rysunkach w artykule, wraz z informacją o ich pochodzeniu. Dodatkowo należy tworzyć sieci cytowań w postaci indeksów cytowań, co robione jest już przez wielu wydawców. Teksty artykułów, podczas wykonywania prac redakcyjnych, powinny być kodowane w XML;

[30] Takimi modelami są na przykład wspomniany już *pay per view* oraz płatności pobierane od autora artykułu.

[31] Adres internetowy: http://www.amazon.com/.

materiał ten należy później udostępniać równolegle z dokumentami w formacie PDF. Zarówno teksty publikacji, jak i surowe dane związane z nimi mogą być udostępniane na warunkach licencji typu Creative Commons[32].

2. Rolą redaktorów jest umieszczenie tekstu artykułu i towarzyszących mu danych w kontekście wiedzy danej dziedziny. Muszą to być więc dobrej klasy specjaliści z zakresu reprezentowanego przez czasopismo. W ten sposób redaktorzy przyczyniają się, łącząc swoje doświadczenie z informacjami dostarczanymi przez autorów, do powstawania nowej wiedzy. Redakcje tworzą i udostępniają m.in. takie narzędzia, jak checkCIF[33], które pozwalają autorom, redaktorom i recenzentom skontrolować poprawność podstawowych elementów tekstu (w przypadku checkCIF jest to poprawność opisanej struktury kryształu) przed jego publikacją. CheckCIF pozwala sprawdzić poprawność struktur opisywanych w tekstach w zakresie krystalografii. System automatycznie przedstawia odpowiednie raporty[34].
3. Uczeni, w procesie publikowania występujący jako autorzy, którzy jednocześnie są użytkownikami publikacji przygotowanych przez kolegów z tej samej dyscypliny, w efekcie najlepiej znają swoją dziedzinę i najlepiej potrafią określić jakość pojawiających się w niej publikacji. Tylko autorzy naprawdę znają powody cytowania tych, a nie innych publikacji w swoich artykułach oraz wyboru rodzaju tych cytowań i cytowanych artykułów. Wydobycie tej wiedzy ukrytej (tacit), na przykład za pomocą narzędzi, takich jak Connotea (adnotacje dotyczące cytowanych publikacji) lub rozszerzonej EndNote[35] Thomsona (aplikacja pozwalająca zarządzać wykazami bibliograficznymi) pozwoli na stworzenie inteligentnych wykazów publikacji. Autorzy są także źródłem surowych, nieprzetworzonych danych, które posiadają w formie czytelnej dla komputera, na przykład jako pliki arkuszy kalkulacyjnych. Jeżeli czasopismo wymaga i umożliwia ich udostępnianie, to będą one przekazywane przez autorów wraz z tekstem do opublikowania w Webie. W ten sposób autorzy zwiększają semantykę swoich tekstów, przydatną podczas ich przyswajania, ułatwiając jednocześnie weryfikację przedstawianych wyników. Uczeni są także twórcami ontologii dziedzinowych i innych wymienionych wcześniej narzędzi, które pozwalają na opanowanie zalewu informacji przez umożliwienie „czytelnictwa strategicznego"[36], nieprzypominającego dawnego czytania

[32] Dostępne pod adresem: http://creativecommons.pl/.

[33] CIF – Crystallographic Information Framework, ontologia struktur krystalograficznych, zawierająca opis około 3300 struktur wraz z podstawowymi atrybutami.

[34] P. Strickland, B. McMahon, J. Helliwell, *Integrating research articles and supporting data in crystallography*, „Learned Publishing" 2008, vol. 11, nr 1, s. 70.

[35] Dostępne pod adresem: http://www.endnote.com/.

[36] A. Renear, C. Palmer, *Strategic reading, ontologies and future of scientific publishing*, „Science" 2009, vol. 325, nr 5942, s. 830.

tekstu od początku do końca, bo polegającego na łączeniu w całość przydatnych urywków z wielu źródeł. Narzędzia te pozwalają uczonym na pracę z wielką liczbą źródeł bez konieczności czytania ich w całości, na co nie ma czasu.

4. Kodowanie informacji w treści artykułu, wykonywane w skali pozwalającej osiągać znaczące efekty, aby było efektywne finansowo (a może w ogóle możliwe), musi być realizowane w sposób zautomatyzowany. Prace dotyczące automatycznego wyszukiwania i kodowania wyrażeń i terminów fachowych przedstawione są w publikacji R. Reis i in[37]. Wynika z nich, że efekty czynności wykonywanych automatycznie, ze względu na niemożność całkowitej eliminacji błędów, powinny być kontrolowane przez ludzi. Narzędzia służące eksploracji tekstu i przetwarzaniu języka naturalnego powinny współpracować z ontologiami dziedzinowymi. Pozwala to na stosowanie języka, który jest także zrozumiały dla komputerów, co daje możliwość komputerowego wspomagania pracy uczonego z publikacją. Szczególne problemy mogą pojawiać się w niektórych dziedzinach, takich jak genetyka, gdzie terminologia nie jest całkowicie jednoznaczna. Zwiększenie efektywności automatycznego rozpoznawania terminów i wiązania ich z ontologiami wymaga dalszych badań w zakresie zarządzania informacją, tworzenia słowników kontrolowanych wspomagających wnioskowanie i technik eksploracji tekstu dla identyfikacji pojęć służących semantycznym adnotacjom. Wielu ekspertów uważa, że w ciągu najbliższych lat dokonany zostanie w tym zakresie znaczny postęp, a za dziesięć lat techniki te stosowane będą rutynowo[38].

4. Zakończenie

Wydaje się pewne, że wydawcy coraz częściej odchodzić będą od prostego kopiowania na nośniki cyfrowe treści swoich wcześniejszych publikacji drukowanych, zmierzając ku publikowaniu semantycznemu. W tym samym kierunku zdąża także rozwój technologii Webu. Należy więc oczekiwać, że przedstawione rozwiązania, choć wciąż eksperymentalne, szybko będą się rozpowszechniać. Tak jak zawsze, również prekursorzy w stosowaniu takiego publikowania zmagają się z problemami (np. brakiem standardów), ale będą także odnosić korzyści, wynikające ze wzrostu zainteresowania ich innowacyjnym produktem, przekładają-

[37] R. Reis et al., *Impact of environment and social gradient on Leptospira infection in urban slums*, „PLoS Neglected Tropical Diseases" 2008, vol. 2 nr 4 (online), dostępne na stronie: http://www.plosntds.org/article/ info:doi%2F10.1371%2Fjournal.pntd.0000228, dostęp: 17 marca 2010 r.

[38] D. Shotton, dz. cyt., s. 90.

cym się na wzrost liczby dostarczanych artykułów, wzrost czytelnictwa i wyższy impact factor. Zastosowanie Webu jako platformy, współpraca w zakresie rozpowszechniania metadanych i cytowań stworzy nowe możliwości powstawania serwisów wspomagających, służących np. wyszukiwaniu informacji. W tym nowym paradygmacie komunikowania naukowego nastąpi znaczna redukcja sztucznych barier między dyscyplinami badań oraz dynamiczna wymiana informacji. W efekcie zwiększy się także współpraca pomiędzy uczonymi z różnych dyscyplin, a to z kolei ułatwi rozwiązywanie kompleksowych problemów nauki.

Użytkownicy Internetu, poszukując informacji pomijają źródła o utrudnionym dostępie, starając się dotrzeć do tych lub podobnych zasobów w jakikolwiek, łatwiejszy sposób. Według Allena Reneara i Carole Palmer, uczeni tylko przeglądają czasopisma dla zdobywania wartościowych informacji, nie czytając ich w całości. Skanują oni teksty w poszukiwaniu przydatnej sobie terminologii, akapitów, diagramów i interesujących ich streszczeń[39]. Przedstawione przykłady wskazują również drogę rozwoju czasopisma elektronicznego i artykułu naukowego, od formatu wspomagającego wydruk (PDF) do semantycznego publikowania, a więc tworzenia wartości dodanej, niemożliwej do uzyskania w publikacji drukowanej. Zmiany w publikowaniu pociągają za sobą zmiany w sposobie korzystania z tekstów naukowych; publikacje stanowią raczej sieć powiązanych informacji, niż zamknięte, dobrze zdefiniowane jednostki, takie jak artykuły z drukowanych czasopism naukowych. Czytanie tekstów jednak nie zaniknie: samo kodowanie nie wystarcza, istotny jest także kontekst, który udostępniany jest przez tekst narracyjny. Ograniczenia w zakresie dalszego rozwoju na tym polu stawia tylko dostępna technologia i wyobraźnia.

Literatura

Berners-Lee T., J. Hendler, O. Lasilla, *The Semantic Web*, „Scientific American" 2001, vol. 284, nr 5, s. 35-43.

Borgman Ch., *Data, disciplines and scholarly publishing*, „Learned Publishing" 2008, vol. 21, nr 1, s. 29-38.

Das S. et al., *Scientific publications on Web 3.0, w: 13th International Conference on Electronic Publishing ElPub 2009*, Milan 2009, s. 107-129.

Evans J., *Electronic publication and the narrowing of science and scholarship*, „Science" 2008, nr 321, s. 395-399.

Nahotko M., *Naukowe czasopisma elektroniczne*, Wyd. SBP, Warszawa 2007.

Renear A., C. Palmer, *Strategic reading, ontologies and future of scientific publishing*, „Science" 2009, vol. 325, nr 5942, s. 828-832.

[39] A. Renear, C. Palmer, dz. cyt., s. 830.

Reis R. et al., *Impact of environment and social gradient on Leptospira infection in urban slums*, „PLoS Neglected Tropical Diseases" 2008, vol. 2 nr 4 (online), dostępne na stronie: http://www. plosntds.org/article/info:doi%2F10.1371%2Fjournal.pntd.0000228, dostęp: 17 marca 2010 r.

Shotton D., *Semantic publishing: the coming revolution in scientific journal publishing*, „Learned Publishing" 2009, vol. 22, nr 2, s. 85-94.

Shotton D. et al., *Adventures in semantic publishing: exemplar semantic enhancements of a research article*, „PLoS Computational Biology" 2009, vol. 5, nr 4 (online), dostępne na stronie: http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000361, dostęp: 17 marca 2010 r.

Strickland P., B. McMahon, J. *Helliwell, Integrating research articles and supporting data in crystallography*, „Learned Publishing" 2008, vol. 11, nr 1, s. 63-72.

Taylor Ch. et al., *Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project*, „Nature Biotechnology" 2008, vol. 26, nr 8, s. 889-896.