Justyna Walkowska
Zespół Bibliotek Cyfrowych, Dział Usług Sieciowych
Poznańskie Centrum Superkomputerowo-Sieciowe

Jeśli nie Web 2.0, to co?

Streszczenie: *Celem artykułu, jest przedstawienie idei semantycznego Internetu (ang. Semantic Web). Jego założenia oraz podstawy technologiczne zostały zarysowane w kontekście instytucji kultury, zwłaszcza tych, które publikują opisy zasobów (czyli metadane, czasem wraz z samymi zasobami w postaci cyfrowej) w sieci. Artykuł powołuje się na znane definicje Web 2.0 i 3.0, starając się odpowiednio umiejscowić względem nich pojęcie semantycznego Internetu. Krótko omawia także aktualnie rozwijane i stosowane ontologie Semantic Web przeznaczone do opisu zasobów dziedzictwa kulturowego, w tym zasobów bibliotecznych.*

Słowa kluczowe: *Web 3.0, Semantic Web (semantyczny Internet), Linked Open Data, biblioteki cyfrowe, internetowe kartoteki autorytatywne*

Kto nie lubi Web 2.0?

Web 2.0, czyli „druga wersja Internetu", to pojęcie, które w mniejszym stopniu jest związane z technologicznym przełomem, a w większym ze sposobem, w jaki użytkownicy korzystają z sieci. W Web 2.0 użytkownik nie jest tylko biernym odbiorcą. Korzystając z Internetu, staje się twórcą: publikuje wiadomości na portalach społecznościowych, ocenia restauracje na ich stronach WWW lub w specjalnie do tego przeznaczonych serwisach oraz udziela się na forach dyskusyjnych, dzieląc się opiniami, ale także konkretną wiedzą.

Określenia Web 2.0 po raz pierwszy użyła Darcy DiNucci już w 1999 r. w artykule[1] adresowanym głównie do projektantów stron internetowych. Nazwała ona ówczesną statyczną postać Internetu „embrionem" tego, co ma dopiero nadejść. Znaczącą rolę w popularyzacji tego terminu odegrała zorganizowana w 2003 r. (przez O'Reilly Media i MediaLive) pierwsza konferencji Web 2.0, na której postulowano paradygmat „Internetu jako platformy". Zakłada on, że aplikacje działają w Internecie (a nie na komputerze użytkownika) w oparciu o dostępne tam dane, tworząc nową wartość i ułatwiając dotarcie do informacji.

Jednym z największych krytyków idei Web 2.0 jest Andrew Keen, autor książki o wiele mówiącym tytule „Kult amatora: jak Internet niszczy kulturę". Oskarża on Web 2.0 o doprowadzenie do zalewu amatorszczyzny w Internecie. Uważa, że Web 3.0 powinien oznaczać powrót ekspertów i przyznanie należnego miejsca ich opiniom w procesie kształtowania Internetu.

Są też tacy, którzy uważają, że „Web 2.0" to tylko chwytliwe hasło, pod którym niewiele się kryje, niezwiązane z żadną faktyczną poprawą jakości. Ukuli oni określenie *Bubble 2.0* (*Bańka 2.0*), sugerujące analogię do ekonomicznej bańki internetowej z lat 1995–2001.

[1] Zob. DINUCCI, D. Fragmented future. *Print* 1999, R. 53, nr 4 (32), s. 220–222.

Web 3.0 — ciągle do przodu

Jeśli się powiedziało „a", trzeba też powiedzieć „b". Skoro zaczęliśmy nadawać numery wersjom Internetu, trudno się dziwić, że co jakiś czas pojawiają się nowe nazwy, definicje i pomysły. Skoro (przynajmniej według niektórych) czas Web 2.0 dobiega końca, to co czeka na nas za rogiem?

Nie istnieje jedna uznana definicja Web 3.0. Prawdopodobnie po raz pierwszy termin ten został użyty przez Johna Markoffa dziennikarza „The New York Times"[2] [6], który zdefiniował go jako rozszerzenie Web 2.0 o mechanizmy związane ze sztuczną inteligencją. Można się spodziewać, że dostępne w Internecie usługi będą „mądrzejsze". Zostaną wzbogacone o możliwość rozumienia języka naturalnego, wnioskowania oraz odkrywania informacji. Podczas korzystania z wyszukiwarek nie będziemy musieli przekopywać się przez dziesiątki zbędnych stron, ponieważ wyszukiwarki zrozumieją nasze intencje i odfiltrują niepożądane wyniki, a pozostałe pogrupują według kategorii.

Web 3.0, według większości publicystów, oznacza personalizację idącą jeszcze dalej niż w Web 2.0. Dużą rolę odgrywają tu smartfony, które pozwalają użytkownikom na natychmiastową publikację informacji o wydarzeniach, w których uczestniczą, przy czym tej informacji często towarzyszą zdjęcia, a także czytelna informacja o lokalizacji.

Jedną z kluczowych idei związanych z Web 3.0 jest semantyczny Internet. Niektórzy publicyści kwestionują jednak gotowość Internetu do wdrożenia postulowanych przez Semantic Web wymagań, inni z kolei mają wątpliwości co do ich przydatności. Wspomniany już A. Keen uważa, że Web 3.0 powinien oznaczać powrót ekspertów i odwrót amatorów generujących treści w Internecie. Jednak oprócz niego i jego zwolenników, większość publicystów i futurologów zgadza się, że Web 3.0 to pójście o krok dalej, a nie rezygnacja z rozwiązań wypracowanych w ramach 2.0.

Semantyczny Internet

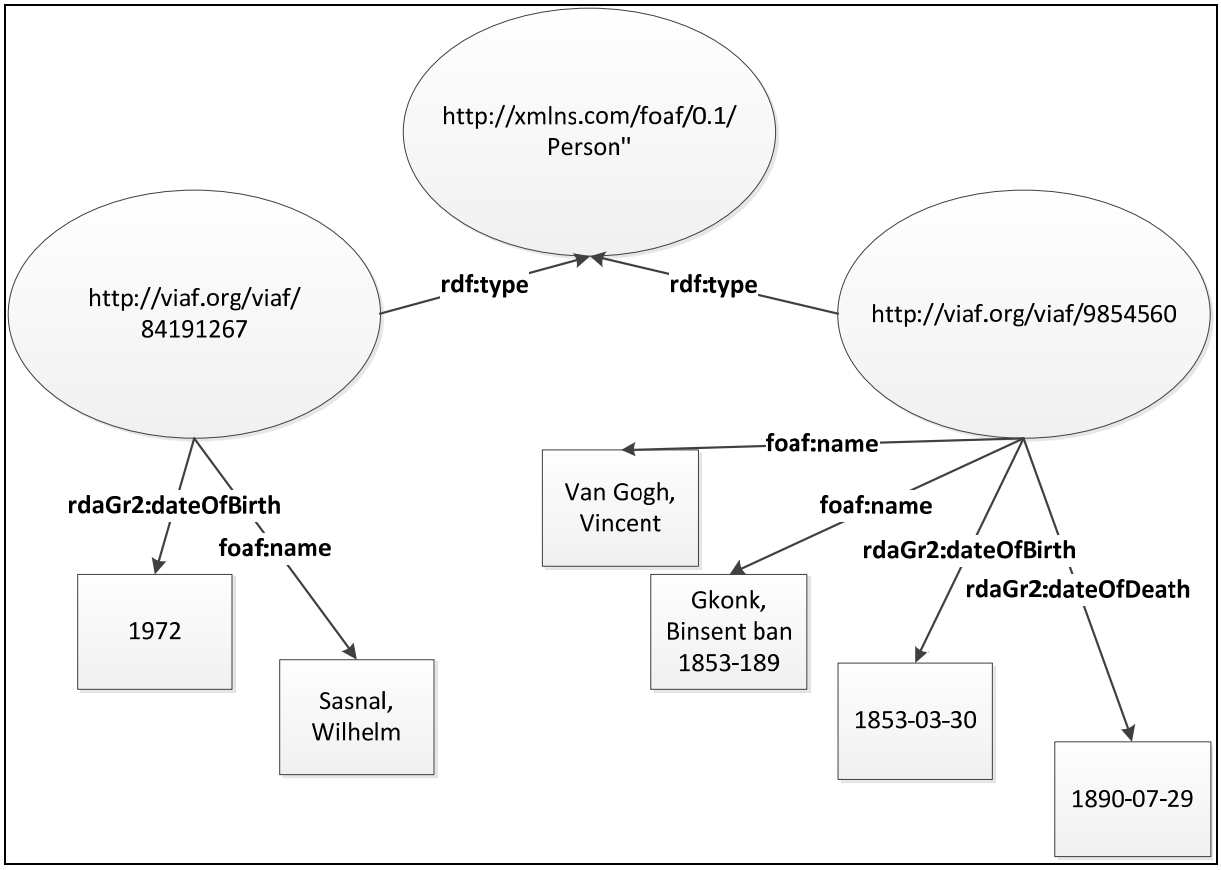
Termin semantyczny Internet[3] bywa stosowany zamiennie z terminem Web 3.0, mimo że Web 3.0, może być interpretowany znacznie szerzej. Semantyczny Internet to międzynarodowa inicjatywa postulująca reprezentowanie danych w formatach umożliwiających ich automatyczne przetwarzanie i integrację, a także automatyczne wnioskowanie w oparciu o nie. Mowa tu nie tylko o nowo tworzonych zasobach, ale również o już dostępnych w sieci danych zaprojektowanych z myślą jedynie o prezentowaniu ich użytkownikom-ludziom. W pewnym uproszczeniu: strony internetowe mają być czytelne zarówno dla ludzi, jak i dla „maszyn", czyli programów, które w różnych celach odwiedzają te same strony.

Głównym przedmiotem zainteresowania są dane, w idealnej sytuacji powiązane między sobą. Pełen nieustrukturyzowanych dokumentów Internet ma zamienić się w „In-

[2] MARKOFF, J. Entrepreneurs see a Web guided by common sense. *The New York Times* [on-line]. November 12, 2006. [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://www.nytimes.com/2006/11/12/business/12web.html.
[3] Zob. HEBELER, J. i in. *Semantic Web programming*. Indianapolis: Wiley Publishing, Inc., 2009.

ternet danych". *Tyle mamy władzy, ile wiedzy* — twierdził Francis Bacon, choć oczywiście, od danych do wiedzy jeszcze długa droga.

U podstaw praktycznej realizacji idei semantycznego Internetu leży format RDF (*Resource Description Framework*). Szczegółowy opis aspektów technicznych nie jest celem tego artykułu, ale warto wspomnieć, że najważniejszą rolę w formacie RDF pełnią *trójki* w postaci: podmiot — orzeczenie — dopełnienie. Podmiotem takiej trójki zawsze jest pewien identyfikowalny obiekt (zasób), orzeczenie to nazwa właściwości tego obiektu, a dopełnienie może być albo analogicznym obiektem, albo *literałem*, czyli wyrażeniem tekstowym. Taka reprezentacja danych prowadzi do powstania grafów, takich jak ten przedstawiony na Rys. 1.



Rys. 1. Graf utworzony przez trójki RDF reprezentujące podstawowe informacje o dwóch osobach. Źródło: rysunek własny (MS Visio)

Owale na Rys. 1. odpowiadają zasobom. Zgodnie z definicją Semantic Web zasób (ang. *resource*) to każdy byt na tyle istotny, żeby otrzymać własny unikalny identyfikator. Na rysunku mamy trzy takie zasoby. Dwa z nich reprezentują osoby: pierwszy (http://viaf.org/viaf/84191267)[4] to identyfikator Wilhelma Sasnala, a drugi (http://viaf.org/viaf/9854560) — Vincenta Van Gogha. Obydwa identyfikatory nadane zostały przez serwis VIAF, o którym za chwilę. Środkowy zasób reprezentuje typ

[4] Wszystkie odesłania do stron internetowych przedstawiają wersję aktualną w dn. 17.02.2012 r.

(http://xmlns.com/foaf/0.1/Person), czyli osobę. Prostokąty odpowiadają literałom. Literały nie mają swoich identyfikatorów, a jedynie zawartość tekstową oraz, opcjonalnie, informację o typie (np. data, liczba) lub języku. W trójce RDF literał może pełnić tylko jedną funkcję — funkcję dopełnienia. Ostatnim elementem rysunku są właściwości, przedstawione jako krawędzie grafu. Łączą one zasoby pomiędzy sobą lub zasoby z literałami, przy czym każda relacja ma określoną semantykę. Zapis z dwukropkiem (*prefiks:nazwa*) to pewien skrót: prefiks to w rzeczywistości określona przestrzeń nazw, w której zdefiniowano właściwości. Celem wprowadzenia przestrzeni nazw jest umożliwienie rozróżnienia relacji o takiej samej nazwie, zdefiniowanych niezależnie od siebie w różnych systemach. Właśnie na wypadek takiej sytuacji, każdy z systemów ma własną unikalną przestrzeń nazw, reprezentowaną zazwyczaj przez adres WWW. Przykładowo *foaf:name* po rozwinięciu ma postać *http://xmlns.com/foaf/0.1/name*. dowiemy się wiele na temat opisywanych bytów. Zdobędziemy wiedzę o tym, że są osobami i poznamy ich nazwiska oraz lata życia. Nie wolno tu przeoczyć pewnego istotnego faktu: zasoby reprezentujące dwóch artystów są ze sobą powiązane poprzez zasób odpowiadający typowi. Dzięki temu, zaczynając od informacji o Van Goghu, odwiedzający stronę wyposażoną w taką reprezentację wiedzy, może łatwo przejść do informacji o innych osobach. Informacje zapisane w formacie RDF mogą być osadzane w treści stron internetowych. Wówczas wersji prezentacyjnej zasobu przeznaczonej dla człowieka (obrazek, tekst w języku naturalnym) towarzyszy informacja, która jest czytelna dla maszyny.

Chcę krótko scharakteryzować jeszcze dwa pojęcia związane z semantycznym Internetem. Dane reprezentowane za pomocą trójek RDF stają się jeszcze lepiej zrozumiałe, gdy towarzyszy im ontologia. *Ontologia* to formalna, jawna specyfikacja wspólnej konceptualizacji[5]. Innymi słowy, jest to specyfikacja typów bytów, które mogą wystąpić w naszych danych oraz właściwości, jakie te byty mogą przejawiać. Tworząc ontologię, definiujemy klasy obiektów (np. „osoba") oraz możliwe właściwości (np. „ma nazwisko", „ma datę urodzin" lub „zna"). Ontologie w środowisku Semantic Web definiuje się najczęściej za pomocą specjalnie do tego celu przeznaczonego języka *Web Ontology Language* (OWL).

Drugie bardzo istotne pojęcie to Linked Open Data (LOD), czyli „powiązane otwarte dane". Otwarte dane to dane publikowane na jednej z otwartych licencji (np. CC0 — http://creativecommons.org/publicdomain/zero/1.0/deed.pl). Dane *powiązane* najczęściej wykorzystują właśnie trójki RDF, a przynajmniej unikalne identyfikatory URI, w sposób, który pozwala na lepszą identyfikację i odnajdywanie zasobów. Duże zbiory danych (np. katalogi biblioteczne, encyklopedie) zawierają odwołania do innych, tworząc tak zwaną „chmurę powiązanych danych" (por. http://lod-cloud.net/).
Tab. 1. przedstawia trzy przykładowe (fikcyjne) fragmenty rekordów metadanych w schemacie Dublin Core opisujących publikacje poświęcone Tadeuszowi Kościuszce. W każdym z przedstawionych trzech przypadków widzimy odwołanie w polu *dc:subject,* zawierającym temat, do Tadeusza Kościuszki. Zastanówmy się jednak, czy maszynie przetwarzającej takie rekordy (na przykład w celu ich prezentacji na stronie instytucji) łatwo jest wywnioskować, że chodzi o tę samą osobę? Do jednej osoby można odwoływać się za pomocą różnych nazw (pseudonimy, wersje nazwi-

[5] GRUBER, T.R. A translation approach to portable ontology specifications. *Knowledge Acquisition* 1993, R. 5, s. 199-220.

ska w innych językach, wersje z inicjałami). Z kolei ta sama nazwa może w niektórych problematycznych przypadkach być dwuznaczna (dwie osoby o tym samym nazwisku i inicjale imienia). Stosując podejście Linked Open Data, możemy zamiast (lub obok) podawania nazwy w najwygodniejszej dla nas wersji, umieścić w metadanych link do którejś z uznanych internetowych kartotek autorytatywnych (w tabeli jest to serwis VIAF), jednoznacznie identyfikując dany byt. W ten sposób poprawiamy nie tylko sposób wyświetlania danych, ale także możliwości ich przeszukiwania.

Tab. 1. Fragmenty trzech rekordów metadanych, opisujących publikacje na temat Tadeusza Kościuszki.

dc:title	Thadée Kosciuszko (1746–1817) : héros de la liberté
dc:subject	Thadée Kosciuszko
dc:title	Tadeusz Kościuszko jego odezwy i raporta. T.5
dc:subject	Kościuszko, Tadeusz — (1746–1817)
dc:title	Uniwersał Połaniecki (z historii sprawy chłopskiej)
dc:subject	*http://viaf.org/viaf/61554270*

Warto mieć na uwadze to, że zanim twórcy aplikacji (w tym serwisów internetowych) odkryją pełen potencjał takiej reprezentacji danych, może upłynąć jeszcze trochę czasu. Dopiero powstają narzędzia umożliwiające wygodne wprowadzanie danych o schemacie będącym ontologią — w tym wypadku nie wystarczy prosty formularz pozwalający pary klucz-wartość; a z kolei ręczne wprowadzenie wszystkich trójek RDF również nie wydaje się najlepszym rozwiązaniem. Dodatkowo, pożądaną funkcjonalnością takiego narzędzia jest możliwość przeszukiwania podczas wprowadzania danych zewnętrznych zbiorów słownictwa i kartotek, dzięki czemu dane od razu byłyby wiązane z ważnymi innymi zbiorami danych. To samo dotyczy kwestii wyświetlania danych użytkownikom. Jednym z prototypowych narzędzi realizujących takie zadania jest system WissKI (http://wiss-ki.eu/).

Semantyczny Internet a instytucje kultury

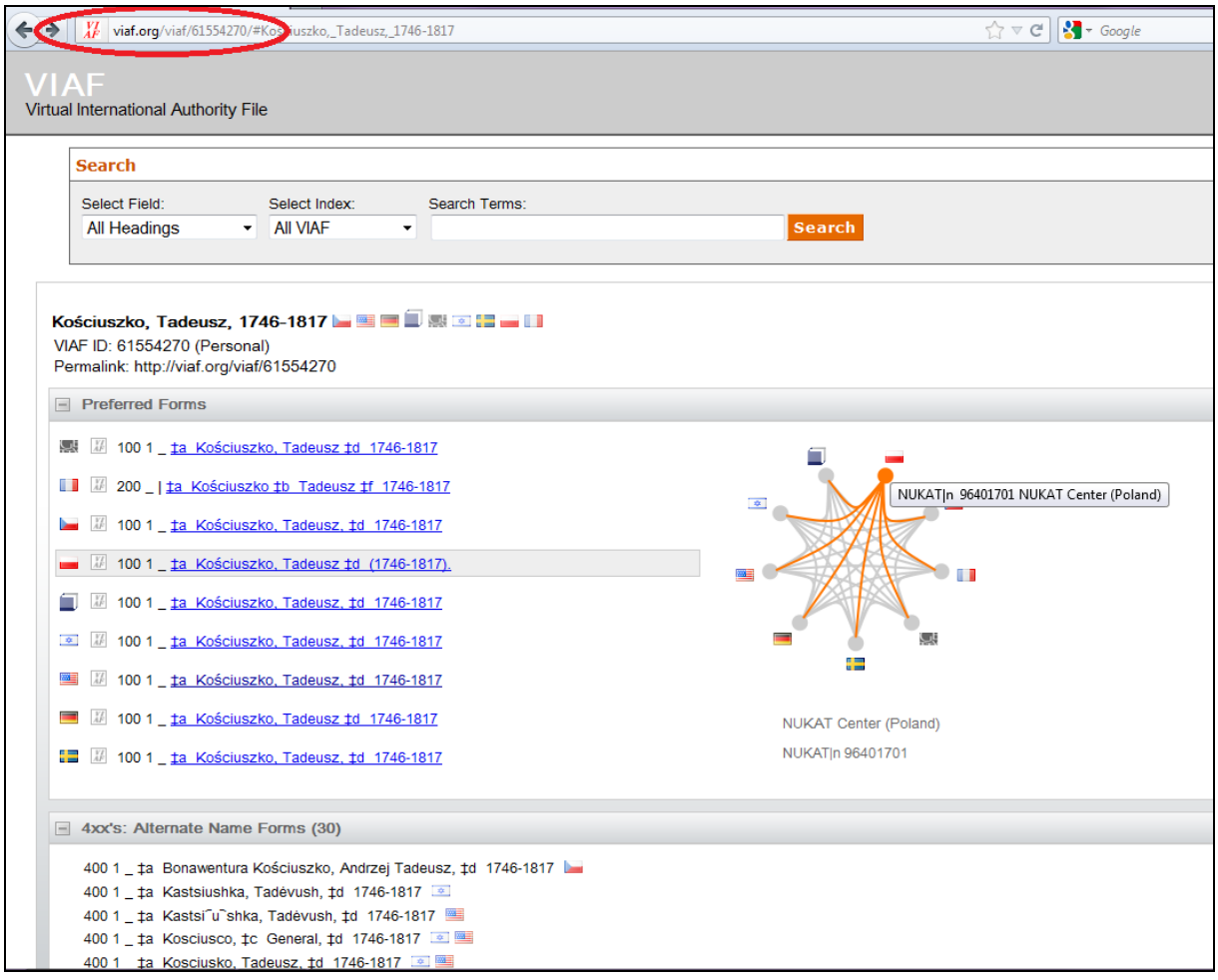
W jaki sposób, jeśli w ogóle, idea semantycznego Internetu przekłada się na świat instytucji kultury? Chciałabym wspomnieć tu dwa kierunki rozwoju. Pierwszy z nich to zgodne z Semantic Web i LOD internetowe kartoteki autorytatywne, drugi to ontologie przeznaczone do opisu danych dziedzictwa kulturowego. W dalszej części przedstawię ontologie CIDOC CRM oraz FRBRoo.

Internetowe kartoteki autorytatywne

Biblioteki i katalogi (nie tylko cyfrowe) dbają o standaryzację między innymi poprzez utrzymywanie kartotek haseł wzorcowych oraz języków haseł przedmiotowych. Prawdopodobnie najpopularniejszym formatem przechowywania tego typu danych jest format MARC 21 — dokładny i stosunkowo jednoznaczny, niestety raczej trudno zrozumiały dla osoby niebędącej z wykształcenia bibliotekarzem (co potwierdzają doświadczenia autorki artykułu), sam w sobie również dosyć trudny do przeszukiwania.

Pewnym krokiem naprzód oraz rozwinięciem idei kartotek wzorcowych w kierunku semantycznego Internetu oraz Linked Open Data są internetowe kartoteki autorytatywne (ang. *authority files*). Nie wzięły się one znikąd — podstawą dużej części z nich są dane pochodzące z bardziej tradycyjnych kartotek. Przykładowo, wspomniany już kilkakrotnie utrzymywany przez OCLC serwis VIAF (*Virtual International Authority File*, http://viaf.org/) udostępnia dane (hasła osobowe, korporatywne oraz tytuły publikacji) pochodzące z kilkunastu instytucji, w tym z polskiego Centrum NUKAT. Każdemu z bytów opisywanych w serwisie nadano unikalny identyfikator URI, poprzez który można odwoływać się do niego w rekordzie metadanych danego obiektu. Dane VIAF są dostępne w kilku formatach, w tym oczywiście w postaci trójek RDF.

Rys. 2. przedstawia stronę serwisu VIAF poświęconą Tadeuszowi Kościuszce. W pasku adresu widzimy (zaznaczony na czerwono) identyfikator tej osoby w serwisie: jest to http://viaf.org/viaf/61554270/. W dalszej części ekranu mamy główne i alternatywne nazwy, wraz z ich źródłami oznaczonymi flagą. Szczególnie zainteresowanych zachęcam do podejrzenia danych w postaci RDF (chociaż w przypadku VIAF są one bardzo proste) — wystarczy po identyfikatorze dopisać „rdf.xml". Zatem dane RDF na temat Kościuszki znajdziemy, wpisując w pasku adresu http://viaf.org/viaf/61554270/rdf.xml.



Rys. 2. Wyniki wyszukania Tadeusza Kościuszki w serwisie VIAF.
Źródło: zrzut ekranu wykonany przez autorkę, serwis viaf.org, przeglądarka Mozilla Firefox

Niektóre spośród internetowych kartotek autorytatywnych są tworzone społecznościowo. Przykładem jest serwis Geonames (http://www.geonames.org/). Obecnie zawiera on informacje (wśród nich: unikalny identyfikator, współrzędne geograficzne, liczbę ludności, typ miejsca itp.) na temat ponad ośmiu milionów miejsc (w tym prawie trzech milionów miejsc zamieszkanych). Jest darmowy i rozwijany przez użytkowników. Ryzykiem związanym z takim modelem tworzenia bazy Geonames jest zwiększona możliwość wystąpienia w niej błędów i przekłamań. Na drugim biegunie znajdują się słowniki (właściwie tezaurusy, czyli słowniki zawierające relacje, a to już z kolei pojęcie bliskie ontologii) oferowane przez amerykański Instytut Badawczy Getty (http://www.getty.edu/research/tools/vocabularies/). Getty oferuje cztery takie bazy, zawierające odpowiednio: terminy z dziedziny sztuki i architektury, nazwy geograficzne, nazwiska twórców oraz dzieła sztuki. Są one starannie tworzone przez ekspertów, ale przez to korzystanie z nich jest płatne. Bardzo ciekawym źródłem są słownik haseł przedmiotowych i kartoteka haseł wzorcowych Biblioteki Kongresu Stanów Zjednoczonych udostępnione nieodpłatnie (również do pobrania jako dane RDF) w roku 2009, po dość długim okresie blokowania tego typu inicjatyw (http://id.loc.gov/).

Ontologia CIDOC CRM

Drugim potencjalnym obszarem wykorzystania idei semantycznego Internetu w instytucjach kultury są ontologie przeznaczone do opisu zasobów dziedzictwa kulturowego. Za ich pomocą złożone dane można porządkować i udostępniać w sieci w sposób zrozumiały zarówno dla ludzi, jak i maszyn.

Jedną z tego typu ontologii, której poświęca się ostatnio wiele uwagi, jest CIDOC Conceptual Reference Model[6] (CIDOC CRM). Jest to ontologia przeznaczona do opisu zasobów związanych z dziedzictwem kulturowym. Prace nad nią rozpoczęto w roku 1996, w tej chwili jest to już standard ISO. Początkowo ontologią zajmowała się grupa Documentation Standards Group wydzielona w ramach Międzynarodowego Komitetu ds. Dokumentacji (CIDOC) przy Międzynarodowej Radzie Muzeów (ICOM). W tej chwili ontologię rozwija i utrzymuje dedykowana grupa CIDOC CRM Special Interest Group.

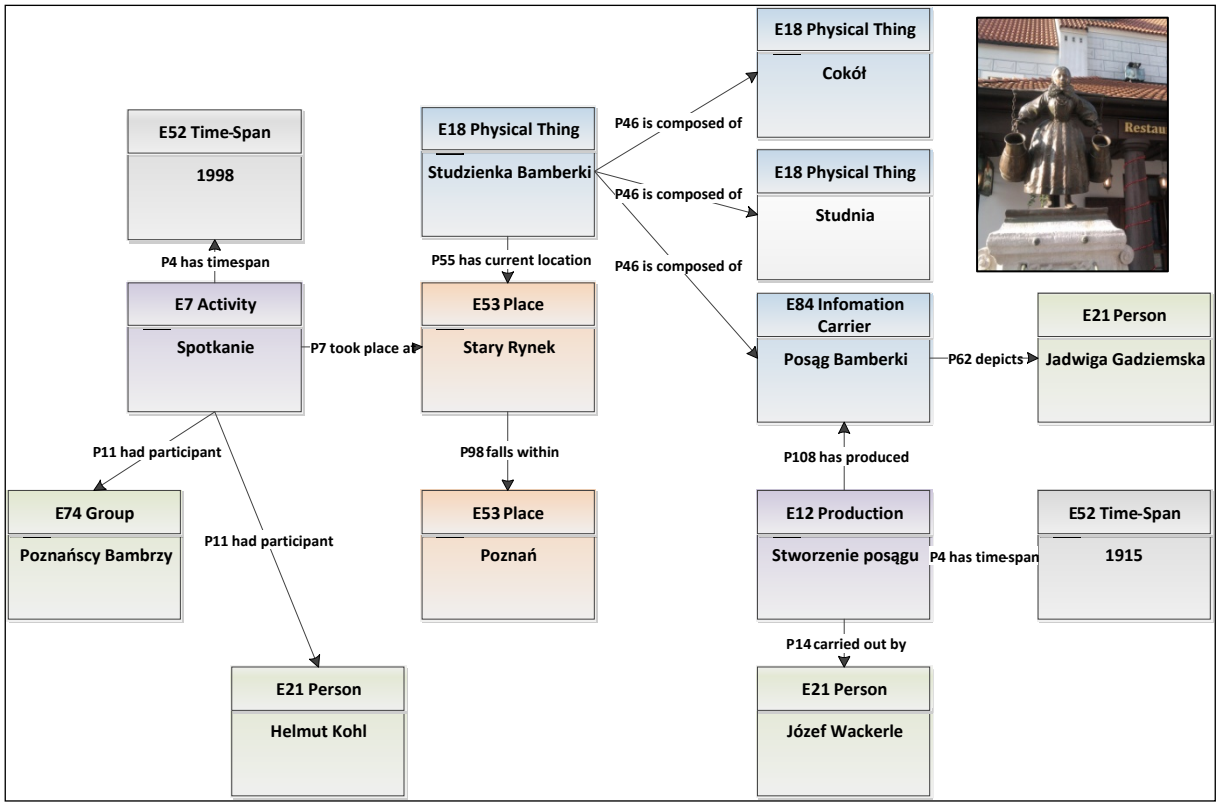
CIDOC CRM definiuje 86 klas i 137 unikalnych właściwości. Należy do grupy ontologii „zdarzeniocentrycznych". Oznacza to, że informacje takie jak data powstania obiektu czy nazwisko jego twórcy nie są dołączane bezpośrednio do obiektu, tylko do odpowiednich, powiązanych z obiektem zdarzeń — w podanym poniżej przykładzie będzie to zdarzenie powstania (wytworzenia) obiektu. Ontologia ta pozwala na wprowadzenie unikalnych identyfikatorów, klasyfikowanie zasobów (można odwołać się do istniejących list typów, na przykład proponowanych przez Getty), reprezentację faktu uczestnictwa osób w zdarzeniach, relacji całość-część pomiędzy obiektami oraz wielu innych relacji.

Rys. 3 przedstawia przykładowy, uproszczony fragment opisu obiektu w ontologii CIDOC CRM. Opisywanym obiektem jest Studzienka Bamberki stojąca na poznań-

[6] CROFTS, N. i in. *Definition of the CIDOC Conceptual Reference Model.* Version 5.0.4 [on-line]. November, 2011 [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf.

skim Starym Rynku. Na rysunku podwójne prostokąty reprezentują obiekty. W górnej części każdego takiego prostokąta podana jest klasa z ontologii, do której należy obiekt. Strzałki odpowiadają relacjom, czyli orzeczeniu w trójce RDF.

Zgodnie z tym, co zostało powiedziane wcześniej, CIDOC CRM jest ontologią zdarzeniocentryczną. W przykładzie widzimy dwa zdarzenia. Jedno z nich jest związane z powstaniem obiektu, a właściwie części obiektu, jaką jest sam posąg Bamberki. Zdarzeniu wytworzenia dzieła przyporządkowany został jego autor, Józef Wackerle, oraz data. Drugie zdarzenie to zdarzenie spotkania kanclerza Niemiec Helmuta Kohla z poznańskimi Bambrami, do którego doszło w 1998 r. na poznańskim rynku, w bezpośredniej bliskości opisywanego obiektu.



Rys. 3. Informacje o Studzience Bamberki zapisane w CIDOC CRM.
Źródło: rysunek własny (MS Visio), fot. Aleksandra Nowak

Dodatkowe informacje, które zostały zawarte w omawianym fragmencie danych, to położenie studzienki (Stary Rynek w Poznaniu) oraz fakt, że składa się ona z trzech elementów: cokołu, studni oraz posągu. Uważny czytelnik zauważy, że posąg został przypisany do innej klasy, niż pozostałe części obiektu — jest to „nośnik informacji" (ang. information carrier), a nie „obiekt fizyczny" (ang. physical thing). Jest to związane z faktem, że pomnik, w przeciwieństwie do pozostałych części, niesie pewną treść informacyjną, związaną z przedstawianą przez niego osobą: Jadwigą Gadziemską, która pozowała rzeźbiarzowi. W ontologii CIDOC CRM *E84 Information Carrier* to podklasa *E18 Physical Thing*, co oznacza, że każdy nośnik informacji jest zarazem obiektem fizycznym.
Osoby zainteresowane praktycznym zastosowaniem ontologii CIDOC CRM mogą odwiedzić stronę projektu CLAROS (http://explore.clarosnet.org/), choć jej interfejs

nie jest jeszcze bardzo rozbudowany. Ontologia ta jest używana również przez Poznańskie Centrum Superkomputerowo-Sieciowe w ramach projektu SYNAT[7], w bazie wiedzy budowanej m.in. na podstawie metadanych z Federacji Bibliotek Cyfrowych, baz danych z Centrum NUKAT czy danych z systemu inwentaryzacji zabytków Muzeum Narodowego w Warszawie[8].

CIDOC CRM dla bibliotek: ontologia FRBRoo

Ontologia CIDOC CRM została stworzona z myślą o reprezentowaniu danych na temat obiektów dziedzictwa kulturowego, do którego zaliczają się oczywiście również obiekty piśmiennictwa. Zasobów przechowywanych w bibliotekach (jak książki) oraz bibliotekach cyfrowych (jak dokumenty w formacie DjVu) nie można rozpatrywać jedynie w kategoriach obiektów fizycznych. Dzieło, które stworzył autor, jest bytem zupełnie innego typu i poziomu niż egzemplarz, który czytelnik może wziąć do ręki.

CIDOC CRM umożliwia reprezentację tej warstwowości: wynik pracy autora można potraktować jako niefizyczny obiekt informacyjny (*E73 Information Object*), zaś egzemplarz jako znany już z wcześniejszego przykładu nośnik informacji (*E84 Information Carrier*). Wówczas jeden obiekt informacyjny (np. *Pan Tadeusz* Adama Mickiewicza) może mieć kilka nośników — w tym rękopis, wydanie książkowe lub wersję na płycie CD.

W niektórych zastosowaniach taka reprezentacja może nie być wystarczająca. Znany większości bibliotekarzy model FRBR[9], zaproponowany przez IFLA (Międzynarodową Federację Stowarzyszeń i Instytucji Bibliotekarskich) zaleca reprezentowanie publikacji aż na czterech poziomach:

1. *Work* (dzieło) to treść intelektualna powstała w umyśle autora, np. *Pan Tadeusz* w „idealnej", „wzorcowej" postaci.
2. *Expression* (realizacja) to intelektualna zawartość danego wydania dzieła, np. tłumaczenie na inny język to także inna realizacja tego samego dzieła.
3. *Manifestation* (materializacja) to zbiór egzemplarzy tego samego wydania, o tej samej grupie cech, jak np. czcionka czy rozmiar kartek lub wydawca.
4. *Item* (egzemplarz) to wreszcie konkretny egzemplarz, który może mieć swoje własne właściwości i swoją historię, np. brak strony, która została wyrwana bądź też odręczną notatkę na marginesie, pozostawioną przez jednego z kolejnych właścicieli.

CIDOC CRM to ontologia dość wysokiego poziomu (dość abstrakcyjna), która nie jest sprzeczna z tym modelem, ale nie definiuje wprost wszystkich klas i właściwości wymaganych przez FRBR. Twórcy ontologii CIDOC CRM zaproponowali jej rozszerze-

[7] Projekt SYNAT jest finansowany przez Narodowe Centrum Badań i Rozwoju (nr umowy: SP/I/1/77065/10). Zob. *SYNAT (System Nauki i Techniki)* [on-line]. [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://sound.eti.pg.gda.pl/synat/.

[8] MAZUREK, C. i in. Transforming a flat metadata schema to a Semantic Web ontology: The Polish Digital Libraries Federation and CIDOC CRM case study. W: *Intelligent tools for building a scientific information platform*. Berlin: Springer, 2012, s 153-177.

[9] *Functional requirements for bibliographic records. Final report* [on-line]. [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.

nie o nazwie FRBRoo (*object-oriented*), które zostało wstępnie zaakceptowane przez IFLA w roku 2008, a obecnie czeka na finalną akceptację. FRBRoo to interpretacja FRBR jako ontologii Semantic Web, przy czym dodane klasy są podklasami tych z CIDOC CRM (np. *Item* to podklasa *Information Carrier*). W ramach wspomnianego wcześniej projektu SYNAT, PCSS nawiązało współpracę z Uniwersytetem w Erlangen, mającą na celu przygotowanie implementacji FRBRoo w języku OWL, aby można było tę ontologię wykorzystywać w szerszej skali w praktyce.

Tylko nie kolejna rewolucja w bibliotekach!

Co to wszystko może oznaczać dla bibliotek? Czy trzeba będzie wyrzucić do kosza bazy opracowane w starszych technologiach? Spokojnie, oczywiście, że nie.

W przypadku opisów obiektów przechowywanych w bibliotekach i muzeach, które same w sobie są bardzo wartościowymi zasobami, należy mówić o ewolucji, a nie rewolucji, chociaż w niektórych momentach ta ewolucja przebiega nieco szybciej niż w innych. Należy zwracać większą niż kiedyś uwagę na jednoznaczność opisów, ponieważ proces ujednoznaczniania jest o wiele trudniejszy dla nienadzorowanej maszyny. Zwiększenie jednoznaczności może wymagać rozszerzenia stosowanego schematu metadanych o dodatkowe pola (w przypadku, gdy jedno pole jest używane do przechowywania kilku różnych typów informacji) oraz wymiany wewnętrznych standardów katalogowania, czasami wraz z koniecznością melioracji, czyli zmiany istniejących opisów. Jeśli dane będą poprawne i uporządkowane, przekształcenie ich do postaci zgodnej z założeniami semantycznego Internetu będzie mogło się odbyć w sposób automatyczny. Niestety ilość danych do konwersji jest tak duża, że niezautomatyzowane rozwiązania tego problemu praktycznie nie wchodzą w grę.

Dostosowanie się do standardów publikacji danych w postaci semantycznej i postaci LOD sprawia, że dane stają się zrozumiałe dla osób (i maszyn) nieznających wewnętrznych formatów stosowanych w danej instytucji. Do danych będzie mogła uzyskać dostęp większa liczba osób. Łatwiejsze stanie się wyszukiwanie (również tak zwane wyszukiwanie fasetowe) i tworzenie odwołań pomiędzy zasobami różnych instytucji. Możliwe są także ciekawsze sposoby prezentacji danych, np. na mapie (w przypadku publikacji związanych z danym miejscem) czy osi czasu.

Ważna jest także możliwość eksploracji, czyli *odkrywania* zasobów. Zdeterminowany użytkownik (lub czytelnik w bibliotece), który dokładnie wie, czego szuka, poradzi sobie przy użyciu dowolnego typu katalogu. Tworząc ustrukturyzowane, poprawne, jednoznaczne i wzajemnie powiązane dane, możemy zainteresować i dłużej zatrzymać czytelnika, który tylko się rozgląda, lub szuka zasobów powiązanych z takim, który już go zainteresował.

Co dalej?

Zgodnie z tym, co napisano na początku artykułu, starając się zdefiniować termin Web 3.0 nie jest on jeszcze ustabilizowany i bywa używany do określenia różnych rzeczy. Większość osób zgadza się, że jego stałym elementem są funkcjonalności związane z semantycznym Internetem.

Odzywają się jednak głosy, że semantyczny Internet w postaci, której rozwój obecnie obserwujemy, to również dopiero początek. W tej chwili semantyczne oznaczanie w Internecie treści, takich jak metadane obiektów dziedzictwa kulturowego, wymaga udziału — lub przynajmniej nadzoru — człowieka. Wraz z rozwojem sztucznej inteligencji, a w szczególności lingwistyki informatycznej (zarówno w kwestii rozumienia języka pisanego, jak i mowy) może się okazać, że stare, niejednoznaczne i niekompletne schematy metadanych, które dzisiaj porządkujemy na potrzeby automatycznego przetwarzania i rozumienia, będą dla maszyny czytelne w stopniu co najmniej takim, jak dzisiaj dla człowieka (który również nie zawsze potrafi od razu odgadnąć, jaki język reprezentuje skrót „mul").

Wpisując w Google hasło „Web 4.0", otrzymujemy ponad 311 000 wyników.

Bibliografia

[1]. CROFTS, N. i in. *Definition of the CIDOC Conceptual Reference Model.* Version 5.0.4 [on-line]. November, 2011 [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf.

[2]. DINUCCI, D. Fragmented future. *Print* 1999, R. 53, nr 4 (32), s. 220–222.

[3]. GRUBER, T.R. A translation approach to portable ontology specifications. *Knowledge Acquisition* 1993, R. 5, s. 199–220. ISSN 1042-8143.

[4]. HEBELER, J. i in. *Semantic Web programming*. Indianapolis: Wiley Publishing, Inc., 2009. ISBN 978-0-470-41801-7.

[5]. *Functional requirements for bibliographic records. Final report* [on-line]. [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://www.ifla.org/files/cataloguing/frbr/frbr_2008.pdf.

[6]. MARKOFF, J. Entrepreneurs see a Web guided by common sense. *The New York Times* [on-line]. November 12, 2006. [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://www.nytimes.com/2006/11/12/business/12web.html. ISSN 0362-4331.

[7]. MAZUREK, C. i in. Transforming a flat metadata schema to a Semantic Web ontology: The Polish Digital Libraries Federation and CIDOC CRM case study. W: *Intelligent tools for building a scientific information platform*. Berlin: Springer, 2012. ISBN 978-3-642-24808-5, s. 153-177.

[8]. SYNAT (System Nauki i Techniki) [on-line]. [Dostęp 17.02.2012]. Dostępny w World Wide Web: http://sound.eti.pg.gda.pl/synat/.

Walkowska, J. Jeśli nie Web 2.0, to co? W: *Biuletyn EBIB* [online] 2012, nr 2 (129), *Koniec 2.0?* [Dostęp: 20.03.2012] Dostępny w World Wide Web: http://www.nowyebib.info/images/stories/numery/129/129_walkowska.pdf. ISSN 1507-7187.