OPENING LIBRARY LINKED DATA TO NATIONAL HERITAGE: PERSPECTIVES ON INTERNATIONAL PRACTICE

Edynburg, Szkocja, 21.09.2012 r.

21 września 2012 r. odbyło się sympozjum *Opening Library Linked Data to National Heritage: Perspectives on International Practice*, zorganizowane przez szkocką grupę ds. katalogowania działającą w ramach Chartered Institute of Library and Information Professionals. Była to druga edycja konferencji Linked Open Data Conference, poświęconej konwersji i publikowaniu danych bibliograficznych i wzorcowych w modelu Linked Data. W tym roku zakres tematyczny wystąpień obejmował metodykę reprezentacji metadanych dla zasobów dziedzictwa kulturowego w kontekście standardów Semantic Web. W sympozjum wzięło udział ponad 60 osób, w tym wiele spoza Wielkiej Brytanii, m.in. z Danii, Finlandii, Hiszpanii, Holandii i Polski. Spotkanie zorganizowano w Edynburgu, w Edinburgh Centre for Carbon Innovation. Program spotkania obejmował wystąpienia siedmiu zaproszonych gości.

Jako pierwszy głos zabrał Richard Wallis z OCLC i zaprezentował referat pt. *Why Link?* R. Wallis jest uznawany za prekursora konwersji danych bibliograficznych do modelu Linked Data. Firma Tallis, której przewodził przez kilka ostatnich lat, jako jedna z pierwszych podmiotów komercyjnych oferowała usługi w zakresie publikowania danych bibliotecznych i wzorcowych w standardach Semantic Web. W kwietniu 2012 r. R. Wallis dołączył do OCLC, co uznaje się za oznakę intensyfikacji prac zmierzających do aktywnego uczestnictwa bibliotek i ośrodków informacji naukowej w rozwoju zasobów sieci trzeciej generacji. Jedną z tez początkowych referatu było podkreślenie roli bibliotek, archiwów i muzeów w rejestrowaniu zasobów dziedzictwa kulturowego oraz znaczenia narzędzi wykorzystywanych do standaryzacji tego procesu (m.in. zasad katalogowania, schematów metadanych, kartotek wzorcowych). Współczesny użytkownik informacji, zdaniem R. Wallisa, zaczyna poszukiwania jednak od wyszukiwarek uniwersalnych, np. Google. Narzędzia te ewoluowały i ich algorytmy procesów przetwarzania nie opierają się już wyłącznie na porównaniu ciągów znaków występujących w zapytaniu oraz w zaindeksowanych zasobach. Jednak zawartość katalogów bibliotecznych czy bibliograficznych baz danych nie jest w pełni indeksowana przez te narzędzia, co powoduje niski poziom ich wyszukiwalności w szerokim kontekście zasobów sieciowych. Jednym ze sposobów rozwiązania tego problemu jest wyrażanie tych zasobów w standardach przetwarzalnych przez aplikacje sieciowe, np. w modelu Linked Data. Polega to na identyfikacji zasobów informacyjnych za pomocą standardu URI, wyrażaniu ich własności oraz relacji z innymi zasobami lub pojęciami, osobami, itd. za pomocą standaryzowanych schematów metadanych dla środowiska sieciowego oraz formułowaniu takich opisów w języku RDF jako standardzie reprezentacji

wiedzy. Wallis nawoływał do zaprzestania kopiowania danych z zewnętrznych zasobów i kartotek wzorcowych do naszych kolekcji. Jego zdaniem, powinno się ustanawiać formalne relacje z tymi zasobami, po to aby nasze kolekcje stały się częścią sieci i były przetwarzalne przez jej aplikacje. Dotyczy to również zaprzestania posługiwania się literałami (ciągami znaków) w odwzorowywaniu cech opisywanego zasobu na rzecz unikalnych identyfikatorów tych wartości, które mogą pochodzić z zewnętrznych źródeł. Tego typu zapis metadanych pozwala na osiągnięcie formalnie powiązanych ze sobą kolekcji i automatyczne pozyskiwanie pożądanych informacji na przykład za pomocą mechanizmów wnioskujących. Wallis przedstawił założenia i realizację projektu OCLC, którego celem było udostępnienie zawartości katalogu World Cat w modelu Linked Data. Głównym założeniem w tym projekcie było przyjęcie mikroformatów jako sposobu zapisu i publikacji metadanych. Efektem dotychczasowych inicjatyw zmierzających do udostępniania danych bibliograficznych i wzorcowych w modelu Linked Data były pliki generowane z relacyjnych baz danych lub specyfikacje zasad komunikowania się zewnętrznych aplikacji z tymi bazami (tzw. Application Programming Interface oraz SPARQL Endpoint). W prototypie World Cat tego typu informacje są zagnieżdżone w strukturze dokumentu HTML lub innego standardu prezentacji informacji w sieci. Zaletą takiego rozwiązania jest możliwość indeksowania takich dokumentów przez wyszukiwarki internetowe. W projekcie tym przyjęto zasoby schema.org jako formalne wykładniki metadanych. Schema.org zawiera specyfikacje metadanych w postaci znaczników, które reprezentują własności typu zasobów informacyjnych oraz osób, miejsc, pojęć itd. Są to proste narzędzia zapisu metadanych, które można osadzić w strukturze dokumentu sieciowego, zwiększając jego wyszukiwalność. Wybór schema.org podyktowany był względami pragmatycznymi, gdyż największe mechanizmy wyszukiwawcze (jak Google, Yahoo, Bing oraz rosyjski Yandex) wykorzystują to narzędzie do strukturyzacji indeksowanych zasobów sieciowych. Prototyp World Cat udostępniono latem 2012 roku i cały czas trwają prace nad jego rozwojem. Wallis przedstawił również założenia grupy roboczej *Schema Bib Extend Community Group*, której celem jest rozszerzenie systemu znaczników schema.org dla zasobów bibliograficznych. Jednym z ostatnich punktów jego wystąpienia był problem licencjonowania metadanych publikowanych w modelu Linked Data i propozycja przyjęcia Open Data Commons jako wyznacznika możliwych rozwiązań w tym zakresie.

Przedmiotem kolejnego wystąpienia był projekt Will's World: Shakespeare Registry Project, realizowany w szkockim Krajowym Centrum Danych – EDINA. Celem projektu jest opracowanie i wdrożenie systemu wyszukiwania informacji o Wiliamie Szekspirze, jego dziełach, twórczości, recepcji, a także o konferencjach naukowych i wydarzeniach innego typu, gdzie poruszane są te zagadnienia. W projekcie przyjęto metodę agregacji jako sposobu zarządzania metadanymi. Na podstawie analizy źródeł opracowano wykaz istniejących baz danych i kolekcji sieciowych, dla których określono zasady komunikacji w kontekście wysyłania zapytań i pobierania metadanych. Testy przeprowadzono na zasobach British Museum opublikowanych w modelu Linked Data. Jedną z form komunikacji z tego typu kolekcjami jest wysyłanie zapytań poprzez tzw. SPARQL Endpoint. Jego formalna specyfikacja zawiera wykaz możliwych składni

zapytań, które mogą być generowane przez zewnętrzne aplikacje w celu pozyskania relewantnych danych. Testy wykazały małą przydatność tego standardu w kontekście prób pozyskiwania informacji za pomocą zapytań o charakterze ogólnym oraz wysoki poziom skomplikowania samej formy generowania zapytań. Projekt ten wpisuje się w obszar zwany *consuming Linked Data*, czyli zagadnień dotyczących wykorzystania istniejących zasobów opublikowanych w tym modelu na potrzeby konkretnych rozwiązań bazodanowych. Wystąpienie miało charakter techniczny i dotyczyło trudności w konstrukcji zapytań do bazy za pomocą SPARQL, jednak poruszono w nim istotny problem funkcjonowania zbyt wielu standardów identyfikatorów dla dzieł, osób, miejsc itd.

Przedstawicielka Wydziału Sztuk Scenicznych Biblioteki Narodowej Francji nawiązała w swoim wystąpieniu do problematyki opisu heterogenicznych zasobów informacji reprezentujących sztuki sceniczne lub ich dotyczących. Pragmatykę opisu wyznacza tutaj standard ISAD(G), lecz dane są przechowywane zarówno w formacie MARC, jak i EAD/XML. Każdy element kolekcji tego wydziału posiada swój unikalny i stały identyfikator wykorzystywany w aplikacji sieciowej zaprojektowanej do przetwarzania i prezentacji metadanych na jego temat. Przedstawiono również problem katalogowania elementów tej kolekcji z uwzględnieniem relacji z innymi elementami zasobów BnF w ramach projektu data.bnf.fr, którego celem jest udostępnienie zasobów tej biblioteki w modelu Linked Data.

Daniel Lewis zaprezentował działalność Open Knowledge Foundation Network (OKFN) na polu otwartych i powiązanych zasobów danych sieciowych. Przedstawił założenia projektu Comprehensive Knowledge Archive Network (ckan.org), którego celem jest udostępnienie metodologii oraz infrastruktury programistycznej i sprzętowej do publikowania otwartych zasobów danych na platformie OKFN. Jednym z narzędzi oferowanych przez OKFN jest serwis http://thedatahub.org/, który pozwala na publikowanie, wyszukiwanie oraz dostęp do ustrukturyzowanych zasobów danych z wielu dziedzin wiedzy. Obszar danych bibliograficznych jest tutaj reprezentowany przez 89 zbiorów danych, które obejmują m.in. kartoteki wzorcowe (np. nazw osobowych Biblioteki Narodowej Niemiec), słowniki języków informacyjnych (np. tezaurus Agrovoc) oraz dane bibliograficzne (np. British National Bibliography). OKFN bierze również udział w projekcie LOD2, współfinansowanym przez Komisję Europejską. Celem projektu jest wypracowanie metodologii oraz rozwiązań programistycznych umożliwiających publikowanie dużych zasobów danych na potrzeby podmiotów związanych z sektorem edukacji, administracji państwowej oraz komercyjnym. Jednym z efektów tego projektu jest serwis http://publicdata.eu/, który pełni funkcję narzędzia agregującego metadane pochodzące z różnych lokalnych zasobów danych o charakterze publicznym.

Kolejne trzy wystąpienia dotyczyły metodyki publikowania danych bibliotek narodowych lub krajowych agencji bibliograficznych w modelu Linked Data oraz problemów, jakie pojawiły się w tych projektach. Były to referaty przedstawicieli Biblioteki Narodowej Szkocji, Duńskiego Centrum Bibliograficznego oraz hiszpańskiego Ontology Engineering Group. W przypadku Biblioteki Narodowej Szkocji ciekawym rozwiązaniem okazało się udostępnienie części kolekcji ikonografii w serwisie społecznościowym Flickr oraz filmów z kolekcji

Scottish Screen Archive w serwisie YouTube. Tego typu inicjatywy spotkały się z dużym zainteresowaniem użytkowników i spowodowały napływ metadanych społecznościowych wzbogacających opisy tych elementów. Projekty szkocki i duński są w początkowej fazie rozwoju i referaty miały formę wykazu otwartych pytań dotyczących efektywnej konwersji i publikowania danych bibliotecznych i wzorcowych w standardach Semantic Web. Charakterystyczną cechą tych dwóch projektów jest zastosowanie podejścia pragmatycznego. Polega to m.in. na przyjmowaniu rozwiązań uznanych za efektywne w innych projektach o tej samej tematyce. Często podkreślanym problemem jest konwersja danych z MARC21 do standardu RDF. Wiąże się to z przyjęciem docelowego modelu konceptualnego i standardów metadanych z wszystkimi logicznymi konsekwencjami związanymi z późniejszym automatycznym przetwarzaniem tych danych. Ważną kwestią w tego rodzaju projektach jest przyjęcie szerokiej perspektywy w publikowaniu danych w modelu Linked Data. Wiąże się to z konwersją kartotek wzorcowych stosowanych do kontroli danych oraz ustanawianiu relacji z zewnętrznymi zasobami danych w sieci. Polega to najczęściej na odwoływaniu się do nazw osobowych z bazy Virtual International Authority File, nazw miejscowych z bazy Geonames oraz osób, pojęć i miejsc do „semantycznej" wersji Wikipedii, czyli DBpedii. Tego rodzaju projekty w Szkocji i Danii zostały wpisane w długoterminowy plan rozwoju bibliografii narodowej. W Hiszpanii, jest to projekt bne.linkeddata.es, który powstał przy współpracy Biblioteki Narodowej Hiszpanii oraz Ontology Engineering Group, która jest grupą skupiającą specjalistów z zakresu ontologii i Semantic Web. Jest to zaawansowany koncepcyjnie i technologicznie projekt badawczy, który, oprócz metodologii, doprowadził również do powstania prototypów aplikacji. Do strukturyzacji danych bibliograficznych wykorzystuje się tutaj model FRBR w jego specyfikacji w RDF. Metadane są formalnie reprezentowane poprzez identyfikatory URI będące wykładnikami własności pochodzących z ISBD oraz RDA. Ciekawą aplikacją opracowaną na potrzeby projektu jest automatyczna kategoryzacja pól i podpól z formatu MARC21 do jednostek z grup 1-3 z modelu FRBR, z wykorzystaniem standardu reprezentacji wiedzy RDF/OWL (http://bne.linkeddata.es/mapping-marc21/).

Jako ostatni wystąpił Gordon Dunsire reprezentujący organizatora sympozjum. Punktem wyjścia jego referatu była potrzeba opracowania i rozpowszechniania tzw. najlepszych praktyk, czyli praktycznych rozwiązań uznanych za efektywne w realizacji projektów wpisujących się w tematykę konferencji. Tym samym dokonał on przeglądu rekomendowanego oprogramowania, które można wykorzystać na poszczególnych etapach oraz scharakteryzował podstawowe sieciowe źródła informacji, z których należy korzystać w tego typu projektach.

Sympozjum Opening Library Linked Data to National Heritage stało na bardzo wysokim poziomie merytorycznym, a prezentowane referaty dotyczyły istotnych problemów pojawiających się w projektach bibliotek mierzących się z koncepcją Semantic Web. Chociaż było to spotkanie jednodniowe, to poruszono wiele ważnych kwestii, pokazując tym samym, jak złożone są to zagadnienia, nie tyle na płaszczyźnie programistycznej czy technologicznej, co koncepcyjnej.

Marcin Roszkowski
Uniwersytet Warszawski, Biblioteka Narodowa