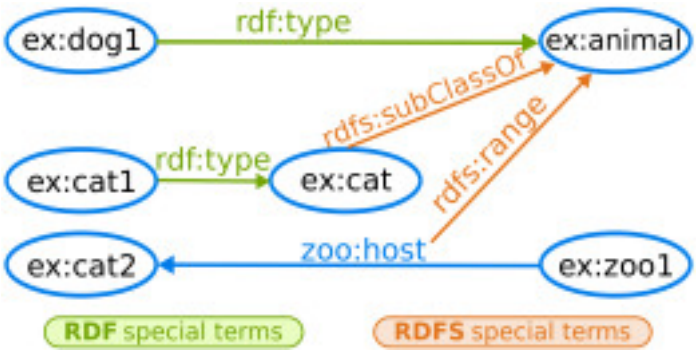
(In English)



Przykład schematu RDF (modyfikacja) - autor Karima Rafes (własne dzieło) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Linked Data to mechanizm którym posługuje się Semantic Web albo “Web 3.0 w budowie”. Te powiązane ze sobą określenia są tak nowe, że nie maja jeszcze ‘oficjalnego’ polskiego tłumaczenia. Na czym polega Semantic Web? Wszyscy używamy World Wide Web (www). Podstawowym składnikiem www sa tak zwane hiperłącza (hiperlink), odnośniki albo odsyłacze do innych stron, źródeł informacji. Kliknięcie w taki odsyłacz (ma w nazwie http) powoduje otwarcie w przeglądarce internetowej nowej strony pozwalającej na rozszerzenie naszej wiedzy lub dalsze zaspokojenie ciekawości. Www została stworzona dla naszej konsumpcji, i jak język naturalny, jest rozumiana przez ludzi.

Jak pisałem poprzednio, komputery są w porównaniu z nami bardzo mało rozgarnięte. Trzeba im wszystko przedstawiać kawa na ławę, metodą łopatologiczną. Ale są za to bardzo szybkie, a przede wszystkim potrafią ogarnąć o wiele więcej danych na raz niż my. A to znaczy, że odszukają w petabajtach informacji to, czego właśnie potrzebujemy. Aby to było możliwe, musimy być dużo bardziej precyzyjni, mieć wiarygodne źródła informacji i system który to wszystko połączy. Tym systemem jest właśnie Linked Data.

Dlaczego interesować się Linked Data? Oczywiście z ciekawości, żeby zrozumieć jak działa dziś świat cyfrowy, który nas otacza; dotyczy to szczególnie archiwistów, bibliotekarzy i innych pracujących w dziedzinie obróbki danych. Jeśli pracujemy w instytucji która posiada jakieś dobrej jakości dane z dowolnej dziedziny, udostępnienie tych danych już teraz w Linked Data podniesie znacząco prestiż tej instytucji na całym świecie.

Najważniejsze zasady Linked Data to używanie odnośników zamiast tekstu (URI) oraz zastosowanie składni prostego zdania twierdzącego: podmiot - orzeczenie - dopełnienie (RDF). Cała reszta to rozwinięcie tych zasad i ich szczegółowa implementacja.

URI

Poprzedno pisałem o unikalnych identyfikatorach . W Linked Data takim identyfikatorem jest URI (Universal Resource Identifier). Zamiast pisać “Aleksander Kowalski ...no..., nie ten poseł na Sejm tylko olimpijczyk, ale nie narciarz tylko hokeista” używamy URI
http://pl.wikipedia.org/wiki/Aleksander_Kowalski_(hokeista)
, który prowadzi nas do jakiegoś autorytatywnego źródła. W przykładzie tym użyłem linku do Wikipedii - encyklopedii dla ludzi; w Linked Data dla komputerów użylibyśmy przykładowo linku do DBPedii,
http://dbpedia.org/page/Aleksander_Kowalski.
Jest tam duża liczba danych posiadających strukturę, takich jak daty, miejsca, kategorie itp, i można zadać bardziej złożone pytanie, np. “podaj wszystkich hokeistów którzy brali udział w Olimpiadzie w Lake Placid razem z Aleksandrem Kowalskim”. Użycie URI zwalnia nas z tłumaczenia za każdym razem, o jaki przedmiot chodzi, a także uniemożliwia lub mocno utrudnia świadome stosowanie niejednoznaczności językowych, ulubione zajęcie polityków.

RDF

Linked Data to idea, w myśl której można używając prostych zdań zapisać informacje o wszystkich obiektach wiedzy ludzkiej - zadanie ogromnie ambitne, ale według twórców Symantic Web wykonalne. Takie proste zdania składają się z podmiotu, orzeczenia i dopełnienia. RDF (Resource Description Framework) jest standardem który określa, w jaki sposób należy tworzyć takie trójczłonowe zdania

Podmiot to obiekt, o którym mówimy w zdaniu. Podmiot musi być określony w postaci URI, t.j. unikalnego identyfikatora. W ten sposób nakierowujemy komputer na jednoznacznie zdefiniowany obiekt. Podmiotem może być każdy obiekt o którym chcemy cos napisać - osoba,

miejsce, książka, strona internetowa itp. W naszym przykładzie będzie to Aleksander Kowalski.

Orzeczenie (często orzeczenie złożone) albo predykat, to druga część zdania. Określa ono jakąś własność, zależność, rodzaj itp. Przykładowo może to być “urodził się dnia…”, “należy do kategorii…”, “ma współrzędne geograficzne…” itp. Orzeczenie też musi mieć charakter URI, aby wskazywać na dobrze zdefiniowany (w języku naturalnym) predykat.

Dopełnienie albo obiekt to wartość, którą w zdaniu przypisujemy podmiotowi. Może to być wartość literalna (np. 1902) ale może też być URI, jeśli wartością jest jakiś obiekt który jest opisany osobno (w tym wypadku nie podajemy jego nazwy, która może być niejednoznaczna, a jego unikalny identyfikator, URI).

W języku naturalnym łatwo takie zdania napisać:

Aleksander Kowalski, ...no..., nie ten poseł na Sejm tylko olimpijczyk, ale nie narciarz tylko hokeista, urodził się siódmego października 1902 roku.

Aleksander Kowalski (ten sam co powyżej) był polskim hokeistą.

Aleksander Kowalski zginął w Katyniu.

W RDF musimy nad tym trochę popracować. Po pierwsze musimy określić URI podmiotu - załóżmy że dla skrócenia zdefiniujemy **ak** jako http://dbpedia.org/page/Aleksander_Kowalski (i dodamy na końcu
/**p**
aby pamiętać, że podmiotem jest osoba a nie strona internetowa) Pierwszym zdaniem będzie informacja dla komputera, że obiektem jest osoba:

ak:p rdf:type foaf:person

Gdzie ak:p to podmiot (mówimy o osobie Aleksandra Kowalskiego), rdf:type to predykat ‘typ’, określony w standardzie RDF, a foaf:person to wartość - ‘osoba’ - klasa określona w standardzie FOAF (o czym dalej). Kolejne zdania to:

ak:p dbpedia-owl:birthDate 1902-10-07

ak:p dc:description “Polish ice hockey player”

ak:p dbpedia-owl:deathPlace dbpedia:Katyn_massacre

W pierwszym zdaniu obiektem jest data zapisana w standardzie ISO (patrz “ Czy umiemy pisać daty? ”) - możemy to uściślić dodając definicję typu, np. tak:

ak:p dbpedia-owl:birthDate 1902-10-07^^ xsd:date

W drugim zdaniu obiektem jest tekst dosłowny (“literal”), a w trzecim jest to URI wskazujące na cały nowy zasób ("resource"), które może mieć swoja listę własności wyrażonych jako zdania RDF, a te z kolei swoje własności i tak dalej. W ten sposób powstaje powoli nowy obiekt który, choć istnieje dopiero w postaci szczątkowej, ma już swoja nazwę: Gigantyczny Globalny Diagram (Giant Global Graph).

Co możemy wyrazić w RDF? Sam standard określa tylko małą liczbę podstawowych pojęć i połączeń, można w nim zdefiniować zasoby (resource), klasy lub kategorie, własności, relacje obiektów do klas, spisy albo listy - uporządkowane i nie, itp. Standard podaje też sposób budowania konstrukcji złożonych (“Jan Kozłowski powiedział że Aleksander Kowalski jest hokeistą”) z prostych zdań trójskładnikowych (w procesie tym, zwanym reifikacją, traktujemy wewnętrzne zdanie jako zasób). Innymi słowy definiuje elementy logiki formalnej którą

posługuje się Linked Data.

Formy zapisu RDF

RDF jest językiem przeznaczonym do zapisu informacji o zasobach. Standard RDF definiuje nie tylko semantykę ale także składnię tego języka, wyrażoną w XML (podawaną czasem jako RDF/XML), co można traktować jako ‘kanoniczną’ reprezentację. Przykład poniżej zawiera tę samą informacje co powyżej w formie XML: pierwsza część <rdf:RDF> to definicje, a <rdf:Description> podaje trzy zdania dotyczace tego samego podmiotu, Aleksandra Kowalskiego.

```
<rdf:RDF
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:dbpedia-owl="http://dbpedia.org/ontology/
     xmlns:dbpedia=”http://dbpedia.org”
     xmlns:dc="http://purl.org/dc/elements/1.1/">
     <rdf:Description rdf:about="http://dbpedia.org/page/Aleksander_Kowalski/p">
       <dbpedia-owl:birthDate>1902-10-07</dbpedia-owl:birthDate>
       <dc:description>Polish ice hockey player</dc:description>
       <dbpedia-owl:deathPlace>dbpedia:Katyn_massacre</dbpedia-owl:deathPlace>
      </rdf:Description>
 </rdf:RDF>
```

Powstało także wiele innych składni (zwanych tu “serializacjami”), które maja być albo łatwiejsze w czytaniu, albo wygodniejsze w zastosowaniu do innych technologii, np www. Istnieje zapis zwany Notation 3 albo N3, który jest drugą ‘oficjalną’ składnią RDF. Zdania ilustrowane w pierwszej części są wyrażone w składni podobnej do N3. Są też składnie o nazwie N-Triples, Turtle, TriG, TriX i inne. Składnia RDFa to metoda umieszczania danych RDF wewnątrz stron www, jako atrybutów hiperlinku. Jeśli to wszystko brzmi trochę jak wieża Babel, to prawdopodobnie tak jest. Każda z tych metod pretenduje do bycia “łatwiejszą” i bardziej intuicyjną, choć i tak te wszystkie dane nie są i zapewne nie będą wprowadzane ręcznie. Zobaczymy które składnie ostaną się próbie czasu.

Rozszerzenia

RDF to tylko początek, trzon składni Linked Data. Potrzebne są bardziej rozbudowane słowniki i narzędzia, jeśli chcemy pretendować do uniwersalnego reprezentowania wiedzy. RDF jest w miarę uniwersalny, co pozwala np. na użycie wewnątrz znanych i popularnych standardów metadanych takich jak Dublin Core (DC) czy Friend-of-a-Friend (FOAF) przeznaczonego do

charakteryzacji danych o ludziach. Sam standard ma już wbudowane rozszerzenie, RDF Schema (RDFS), z definicją klas dla budowania ontologii. Powstają całe kompletne języki ontologii takie jak OWL albo SKOS, języki kwerend baz danych "trójkowych' jak SPARQL i inne narzędzia. Są one bardzo ciekawe, ale to jest temat na osobny artykuł.

Gdzie są dane?

Wszystko to bardzo dobrze, można sobie powiedzieć, ale skąd mamy wziąć te URI? Gdzie są dane, co do których możemy mieć zaufanie, że są prawdziwe i nie wprowadzą komputera w błąd? Przecież każdy taki błąd może się rozprzestrzenić jak pożar na inne węzły Semantic Web.

Jest wiele zasobów które są udostępnione w Linked Data. Liczba wiarygodnych źródeł danych rośnie wykładniczo. Dostępne są dane geograficzne i dane o ludziach - szczególnie autorach, zbierane przez biblioteki na świecie. Dużą nadzieję pokłada się w dostępie do danych naukowych, w szczególności w takich dziedzinach jak genetyka, meteorologia czy fizyka gdzie zasoby danych liczą się w petabajtach. DBPedia jest odzwierciedleniem Wikipedii ale z danymi które mają strukturę; instytucje takie jak Library of Congress albo New York Times udostępniają coraz więcej informacji. Można to zobaczyć juz dziś, wpisując na przykład w przeszukiwarkę Google zdanie "Birth date of Jozef Pisudski" (działa to na razie tylko po angielsku). Dostaniemy już nie tylko linki do stron, ale konkretną odpowiedź (5 grudnia 1987) plus, jako bonus, obszar na stronie z innymi danymi encyklopedycznymi o Marszałku. Omówienie źródeł danych, których jest ogromna ilość, to jednak temat na następny blog.

Czytaj więcej

- Strona Linked Data
- Linked Data w Wikipedii
- Semantic Web w Wikipedii
- Co to jest RDF i czym to sie je (po angielsku, jak i inne zródła powyżej)

Marek Zieliński

Artykuł ukazał się 25 listopada 2013 w *Blogu archiwistów i bibliotekarzy Instytutu Piłsudskiego*

Może Cię też zainteresować

- Wstęp do standardów metadanych
- Koperty na zdjęcia cyfrowe
- Digital Humanities