W zamierzchłych czasach, kiedy żyliśmy w wioskach, kontekst wystarczał do rozwikłania niejednoznaczności w języku, a w szczególności homonimów. Słowa takie jak zamek, bałwan, para, bal albo rakieta muszą być użyte w kontekscie aby były zrozumiałe (jaki obiekt przychodzi Ci na myśl, kiedy głośno wypowiesz słowo “zamek”?). Globalizacja informacji, szczególnie po powstaniu Internetu, wymaga szczególnej staranności w definiowaniu kontekstu. W powieści “Tajemniczy ogród” Frances Hodgson Burnett, występuje ptaszek, “robin”, albo po polsku rudzik. W USA “robin”, to zupełnie inny ptak, drozd wędrowny i wszyscy czytelnicy tej książki w Stanach Zjednoczonych są wprowadzeni w błąd. Użycie słowa “football” na określenie zupełnie różnych sportów w różnych częściach świata prowadzić może tylko do śmiesznych nieporozumień, użycie słowa bilion, które czasem znaczy tysiąc milionów, a czasem milion milionów (zależnie do miejsca i czasu) może prowadzić do poważnych już konsekwencji, szczególnie finansowych.

W języku naturalnym niejednoznaczność jest przyprawą, smakiem - bez niej nie było by insynuacji, niedopowiedzeń, podtekstów, poezji. Ale w naukach, zarówno ścisłych jak i humanistycznych, niejednoznaczność jest trucizną wiedzy, i musi być bardzo starannie unikana. W roku 1735 Karol Linneusz opublikował “Systema Naturae”, pierwszą sytematyczną próbę wprowadzenia jednolitego nazewnictwa w biologii. W roku 1782 Louis-Bernard Guyton de Morveau opublikował rekomendacje jednolitego nazewnictwa chemicznego. Obie te publikacje były tylko początkiem bardzo złożonych (i ciągle ulepszanych), działajacych obecnie systemów nazewnictwa w biologii i w chemii. Podstawowym językiem - lingua franca - tych systemów jest łacina w przypadku klasyfikacji biologicznej i angielski w chemii, a nazwy w innych językach sa tylko (mniej lub bardziej jednoznacznymi) tłumaczeniami.

Jakie więc mamy możliwości przypisania jakiemuś obiektowi jednoznacznej nazwy? Jedną z nich jest **umieszczenie całej informacji w nazwie**. Zaletą tej metody jest to, że nazwa jest czytelna dla czowieka, więc mozna stosować ją w tekstach. Oznacza to, że często musimy rozszerzać nazwy o dalsze ich składniki tak aby uzyskac jednoznaczność, a więc rudzik będzie się teraz nazywać European Robin, a drozd wędrowny otrzyma miano American Robin. Tą drogą poszło np. nazewnictwo chemiczne, tworząc nazwy od prostych do takich jak (3S,6R,7E,9R,10R,12R,14S,15E,17E,19E,21S,23S,26R,27R,34aS)-9,10,12,13,14,21,22,23,24,25,26,27,32,33,34,34a-hexadecahydro-9,27-dihydroxy-3-[(1R)-2-[(1S,3R,4R)-4-hydroxy-3-methoxycyclohexyl]-1-methylethyl]-10,21-dimethoxy-6,8,12,14,20,26-hexamethyl-23,27-epoxy-3H-pyrido[2,1-c][1,4]-oxaazacyclohentriacontine-1,5,11,28,29(4H,6H,31H)-pentone (lek używany przy przeszczepach - jak widać przy pewnym poziomie komplikacji nazwa przestaje

jednak być czytelna). Podobna drogą poszło bibliotekarstwo, tworząc systemy kontroli autorytatywnej, o czym dalej.

Drugą mozliwością jest stworzenie **rejestru**. W tym wypadku potrzebna jest instytucja ciesząca się autorytetem, o długim spodziewanym czasie życia. Instytucja taka organizuje i publikuje rejestr, w którym każdemu obiektowi przypisuje unikalny identyfikator - tekstowy lub numeryczny. Ten identyfikator nie może już zastąpić nazwy, ale może pozwolić na odróżnienie obiektu od innego, o podobnej nazwie. W chemii jak i w bibliotekarstwie istnieja takie rejestry. W chemii numery Chemical Abstract Service (CAS) są przypisane pierwiastkom, związkom lub mieszaninom; Bliblioteka Kongresu (Library of Congress) nadaje unikalne numery książkom i czasopismom jako ISBN i ISSN. W ogólności system ten polega na zastąpieniu nazwy parą: klucz-etykieta. Klucz jest unikalnym identyfikatorem, etykieta jest czytelnym dla czlowieka opisem (w pierwszej metodzie etykieta była jednocześnie kluczem). Oczywiście klucz może teraz wskazywac na cały obiekt, nie tylko jego etykietę. Numer ISBN wskazuje na konkretne wydanie książki, a tytuł jest tylko jedna z wielu etykiet (charakterystyk) tej książki.

Trzecim, najnowszym i najbardziej nowoczesnym rozwiązaniem jest użycie **Linked Data**, konceptu Sieci Semantycznej (Semantic Web) propagowanej przez twórcę WWW Tima Bernersa-Lee. W sieci jest już ogromna liczba zasobów, które możnaby użyć, gdyby tylko istniała możliwość automatycznego ich wybierania i linkowania. Takie możliwości powstają w szybkim tempie, i coraz więcej instytucji bierze w tym systemie udział. Poniższy diagram jest fragmentem mapy obrazującej połaczenia
Linked Data
- każdy węzeł sieci to jakaś organizacja publikująca otwarte dane dostępne w tym systemie.

Na czym ten system polega? W ogólności, zamiast kopiowania informacji używa się *linku* do jej źródła. Jeśli więc piszemy o ptaszku “robin”, zamiast wyjasniać o którego ptaszka chodzi, możemy dać w tekście elektronicznym link, który doprowadzi nas do źródła informacji. Ale to jest tylko początek. Mając dane połaczone można np. napisać zdanie: “populacja Łodzi wynosi teraz …. osób”. Ukryty w tym miejscu link wykona szybka kwerendę i poda najnowsze dane z autorytatywnego źródła. Możliwości są ogromne, wymagane jest tylko, aby instytucje posiadająca dane udostępniła je w sposób otwarty, tj. dostępny dla każdego w sieci. Przykładowo DBpedia używa algorytmów i sztucznej inteligencji aby wyekstrahować w Wikipedii dane które mozna podawać w sposób zautomatyzowany innym witrynom które takiej informacji poszukują.

Biblioteki i archiwa przerabiają i udostępniają ogromne ilości informacji i unikalne identyfikatory ogromnie zwiększaja tej informacji precyzję i niezawodność. Linked Data jest bez wątpienia

przyszłością, ale aby móc używać tego systemu potrzebne jest właściwe oprogramowanie. W międzyczasie powinnismy się opierać na dwóch pierwszych metodach. Miejsca, osoby, instytucje, publikacje, zasoby archiwalne, hasła tematyczne to przykłady obiektów które mogłyby skorzystać z unikalnych indentyfikatorów. Hasła takie nie tylko występują często w wyszkiwaniu danych (search), ale są podstawa tworzenia grup i kategorii w przeglądaniu danych (browse).

Nazwy geograficzne zmieniają się w czasie, miasta o tej samej nazwie występuja w wielu miejscach (w USA jest 10 miast o nazwie Warszawa), a w różnych językach maja różne nazwy. Przy podawaniu miejsc geograficznych przydatny jest serwis GeoNames , największa, publicznie dostepna baza geo-danych, zawierająca miliony rekordów. GeoNames jest również właczona w system Linked Data. Wpisanie nazwy miejsca (np. Łódź) otwiera stronę z ogromna liczba informacji na temat tego miejsca, pokazuje alterntywne nazwy (w tym historyczne i w obcych językach), grupuje hasła Wikipedii dotyczące tego miejsca (i okolic) itp. Współrzędne geograficzne są oczywiście przydatne, gdyż określają jednoznacznie miejsce na Ziemi, ale nie wystarczające, gdyż ten sam punkt może należeć do wielu obiektów, np. Nowy Jork jest nazwą stanu, powiatu, miasta, dzielnicy, itp. i współrzędne Zamku Belwederskiego w Central Parku mogą odpowiadac każdej z tych jednostek geograficznych (i wielu innym). Drugim cennym zasobem przy identifikowaniu miejsc jest Wikipedia, która podaje encyklopedycznie zebrane informacje na temat danego miejsca. Link do GeoNames lub Wikipedii daje dobre przybliżenie unikalnego identyfikatora.

Instytucje pojawiają się i znikaja, zmieniaja nazwy, łaczą się i dzielą. Poszczególne kraje posiadają rejestry firm których założenie wymaga rejestracji, ale spisy te są trudno dostępne i ograniczone. W niektórych krajach (USA, Kanada, Niemcy i Wielka Brytania) istnieja rejestry organizacji , ograniczone głownie do bibliotek, muzeów, archiwów i tym podobnych, które pozwalają na uzyskanie unikalnego identyfikatora danej instytucji. Rejestr Biblioteki Kongresu USA nie jest ograniczony tylko do tego kraju, ale istnieje w nim minimalna (12) liczba organizacji z Polski. (Instytut Piłsudskiego w Ameryce posiada taki identyfikator: US-NyNyJPI). Użycie takiego identyfikatora powoduje na przykład, że każdy archiwalny rekord EAD Instytutu, oprócz danych dotyczących zasobu (kolekcji, zespołu archiwalnego), posiada również identyfikator jednoznacznie okreslający repozytorium tego zasobu.

Biblioteki posiadają dużą tradycję w tworzeniu słowników unikalnych identyfikatorów, zarówno w klasyfikowaniu zasobów kultury (systemy klasyfikacji Deweya, dziesiętnej, Biblioteki Kongresu itp.), jak i w tworzenu spisów osób, haseł tematycznych itp. Aby zostać dodanym do spisu osób trzeba zwykle być autorem, współautorem lub tłumaczem książki, więc są to spisy dość ograniczone. Hasła wzorcowe podaja wzorcowy sposób definiowania tematu, podobnie jak

klasyfikacja. Używanie wzorcowego hasła tematycznego ułatwia zawsze grupowanie obiektów (choć niekoniecznie ich odszukanie). Hasła wzorcowe Biblioteki Narodowej są dostępne w sieci. Korzystanie z nich i katalogów nie jest łatwe, jak mozna się przekonać próbując znaleźć hasła charakteryzujące czasopismo Instytutu Piłsudskiego “Niepodległośc” wydawane od 1929 roku.

Cennym zasobem haseł wzorcowych, działajacym także w systemie Linked data jest VIAF , baza danych łacząca słowniki haseł wzorcowych z USA i wielu innych krajów. Nie tylko stanowi ona pjedyńcze żródło danych, ale uwzględnia wersje alternatywne i wersje w wielu językach. Przykładowo, wybierając hasło
Józef Piłsudski
otrzymujemy w zasadzie pojedyńczy rekord, z odsyłaczami do 12 baz danych w różnych krajach. Próba zlokalizowania naszego Instytutu nie spotyka sie już z takim sukcesem: znajdujemy co najmniej cztery różne rekordy (każdy z innym identyfikatorem) dotyczące tej samej instytucji:
http://viaf.org/viaf/278200980
,
http://viaf.org/viaf/151002901
,
http://viaf.org/viaf/262858213
,
http://viaf.org/viaf/277221969
. (Pomijam tu hasła dotyczące Instytutu Piłsudskiego w Londynie, czy też Instytutu Piłsudskiego w dwóch lokalizacjach na raz: Londynie i Nowym Jorku). Widać więc, ze dyskryminacja i usuwanie duplikatów nie jest najmocniejsza stroną takiego katalogu.

W tym miejscu ponownie widać wartość Wikipedii jako źródła informacji. Hasła nie moga się duplikować, a hasła o identycznym lub podobnym brzmieniu są podawane na stronach ujednoznaczniających. Zarówno “Niepodległość” jak i Instytut Piłsudskiego są do znalezienie i mają całkiem unikalne odnośniki. Dzięki pracy tysięcy wolontariuszy istnieje więc system który posiada unikalne hasła, ich opisy (od niezłych po doskonałe) i odsyłacze do haseł w dużej liczbie języków. Przy 4 milionach haseł w jezyku angielskim i milionie po polsku, jest to potężne i wiarygodne żródło którego można (i warto) używać.

Więcej o unikalnych identyfikatorach

- Dowcipny filmik o unikalnych identfikatorach (po angielsku)
- Hasła wzorcowe w Wikipedii (po angielsku)

Marek Zieliński

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

Może Cię też zainteresować:

- Wstęp do Linked Data
- Wstęp do standardów metadanych
- Digital Humanities