Część 1

Przy planowaniu procedur i etapów pracy projektu digitalizacji zasobów archiwalnych Instytutu zastanawialiśmy się nad tym, jakiego standardu użyć przy opisie digitalizowanych dokumentów. Po wykonaniu skanu czyli zapisu obrazu dokumentu, niezbędnym jest jego opisanie tak, aby możliwe było znalezienie interesującej czytelnika informacji. Idealnie byłoby dokonać transkrypcji całości materiału, ale przy dużej ilości ręcznie pisanych dokumentów były to tylko marzenia. Proces znajdywania i organizowania informacji o tekście, zdjęciu itp. czyli tak zwanych metadanych (danych o danych) oraz zapisywania ich w odpowiedniej bazie danych jest najważniejszyma (i najbardziej czasochłonnym) etapem digitalzacji.

Jest wiele schematów i standardów zapisu i transportu metadanych, więcej niż można łatwo ogarnąć ciekawie brzmiących skrótów: DC, EAD, MARC, MODS, TEI, AACR2, CCO, CDWA, DACS, FOAF, ISAD(G), METS, OAI-PMH, OAIS, OWL, POWDER, PREMIS, RDA, RDF, SWORD itp. Nawet ograniczając się do standartów opisu metadanych zawartości dokumentów, mieliśmy do wyboru DC, EAD, MARC i TEI . Przy testach pojawił się problem hierarchizacji informacji, który w zasadzie ciągle jest z nami, mimo prób jego oswojenia.

Człowiek ma tendencję, jak się domyślam wrodzoną, do grupowania i hierarchizowania pojęć. Wiąże się to ściśle z intuicyjnie ograniczoną wielkością zbiorów które można myślą ogarnąć. Dlatego też rok dzieli się na miesiące i tygodnie, doba ma tylko 24 godziny (w krajach anglosaskich dzień ma dwie częsci po 12 godzin, AM i PM), godzina ma 4 kwadranse po 15 minut itp). Hierarchia albo taksonomia polega na łączeniu pojęć w grupy, a tych grup w większe grupy tak, że każdy obiekt umieszczony w ściśle jednym miejscu w hierarchii. Człowiek jako gatunek mieści się w hierarchii zapoczątkowanej przez Linneusza: 1. domena: Eukarionty, 2. królestwo: Zwierzęta, 3. podkrólestwo: Tkankowce właściwe, i tak dalej aż do 21: gatunek: Człowiek rozumny. Ta hierarchia jest ścisła, gdyż każdy z nas ma dokładnie dwoje rodziców, ale wiekszosć hierarchii pojęć jest tylko przybliżona. Łaczenie pojęć w grupy jest przydatne, gdyż można wtedy czerpać własności obiektu z opisu grupy, stosując inferencję. Jeśli więc mamy przedmiot z grupy przybory biurowe, podgrupa narzędzia do pisania, pod-podgrupa przyrządy zawierające grafit, to z małym prawdopodobieństwem błędu wyobrazimy siebie ołówek. Grupowanie ludzi pod względem ich pogladów politycznych jest ulubionym zajęciem większości, mamy więc grupy: liberałów i konwerwę, komunistów (a nawet komuchów) i zatwardziałych prawicowców, itp. Po przypisaniu polityka lub kolegi do jednej z takich grup nie

trzeba się już zastanawiać nad jego indywidualnymi zaletami czy umiejętnościami, wystarczy użyć cech całej grupy :-)

Prawie każdy rodzaj informacji można traktować jako hierarchiczny. Prosty adres Instytutu Piłsudskiego, *180 Second Avenue, New York, NY* można traktować jako hierarchię: galaktyka: Droga Mleczna, gwiazda: Sol, planeta: Ziemia, kontynent: Ameryka Północna, państwo: USA, stan: Nowy Jork, miasto: Nowy Jork, dzielnica: Nowy Jork, ulica: Second Avenue, numer ulicy: 180, piętro: drugie. Ten przykład ilustruje dobrze problemy zbytniej hierarchizacji. Po pierwsze wiele poziomów jest domyślnych, i nie ma potrzeby ich specyfikowania. Po drugie nie wszystkie poziomy musza być znane przy opisie obiektu, a brak jednego zatrzymuje cały proces. Warsaw jest miastem w stanie Nowy Jork i w 10 innych stanach USA (a także jest stolicą Polski). Jeśli nie wiemy o którą Warszawę chodzi, nie możemy informacji zapisac w ogóle w systemie hierarchicznym, jakkolwiek możemy w płaskim.

Czym więc jest system płaski? Jest to system zapisu informacji który świadomie odrzuca (albo mocno ogranicza) hierarchiczność, idzie więc nieco pod prąd instynktowi. Dane przypisane do danego obiektu (metadane) posiadają tylko jeden, najwyżej dwa poziomy. Tracimy więc głębokość hierarchii, ale także jej sztywnośc, zyskujemy postotę w użyciu. Prostota jest bardzo ważna. Ogromna liczba ludzi na świecie przetwarza informację, tworzy zapisy elektroniczne. System hierarchiczny działa dobrze tylko wtedy, jeśli jest dobrze opanowany, wbity w pamięć. Kto z czytelników potrafi wymienić wszystkie 21 poziomów klasyfikacji Homo Sapiens? Prosty system opisu danych, dla użycia którego nie trzeba odbywać studiów, na szanse na to, że będzie powszechnie używany. System w którym niepełne lub niedokładnie znane informacje mogą być dodane do opisu obiektu jest szczególnie przydatny w sytuacjach, gdzie informacja bywa niekompletna, jak np. w genealogii czy w archiwistyce.

W następnym odcinku będzie więcej o płaskich i hierarchicznych standardach metadanych i o naszych Instytutowych wyborach.

Marek Zieliński

Artykuł ukazał się 20 lipca 2012 w *Blogu archiwistów i bibliotekarzy Instytutu Piłsudskiego*

Może Cię też zainteresować:

- Czy umiemy pisać daty?
- Standardy metadanych: EAD