Przy omawianiu zderzenia nauk bibliotecznych, archiwistycznych itp. z komputeryzacją i Internetem, centralne miejsce zajmują metadane i sposób ich wyrażania. Metadane to dane o danych, opisy, wyciągi, oznakowania, indeksy, katalogi itp. Ten artykuł jest wstępem do dyskusji i omówienia różnych aspektów metadanych i ich zastosowań.

Dlaczego w ogóle potrzebne są nam metadane? Najprostszą odpowiedzią jest “dlatego, że komputery są raczej nierozgarnięte”. Niech nas nie zmyli fakt, że umieją grać w szachy lepiej od ludzi - to jest zadanie względnie proste w porównaniem ze zrozumieniem języka naturalnego. Ale mimo tego, że są nierozgarnięte, komputery są w stanie przetworzyć dużo więcej informacji w dużo krótszym czasie niż mózg człowieka, więc jest w naszym interesie tłumaczyć zdania języka naturalnego na język zrozumiały przez komputery.

Krótka wymiana zdań, zrozumiała przez rozmówców, np. “Jurek, znajomy mojego wuja, urodził się w Bielsku w latach pięćdziesiątych” jest dla komputera zupełnie nie do odcyfrowania. Nie jest w stanie zidentyfikować Jurka, mnie, mojego wuja, Bielska ani też daty. Aby to było możliwe, trzeba biedakowi trochę pomóc.

Zanim przejdę do szczegółów, krótkie wprowadzenie dwóch pojęć, które będą nam dalej przydatne: **składnia** albo syntaktyka, i **semantyka** (powiązana z ontologią). Składnia to zestaw reguł (w miarę możności ściśle zdefiniowanych) jakimi posługuje się jakiś język. Semantyka (a w jej rozszerzeniu ontologia) dodaje ‘znaczenie’ - zajmuje się nie tym, jak zdanie jest zbudowane, ale tym, co ono znaczy. Języki naturalne posiadają obie cechy, często mocno ze sobą splątane. W konstrukcji języków komputerowych te dwa elementy są zwykle łatwe do rozdzielenia. Na przykład zdanie z języka BASIC

```
FOR i = 1 to 12 STEP 2
PRINT i
NEXT
```

ma bardzo ściśle określoną składnię z pewnymi opcjami, np. STEP 2 można opuścić, ale NEXT jest wymagane. Znaczenie, czyli semantyka, to działanie tego programu w komputerze, który dokona pewnych operacji: przyjmie dla zmiennej i wartość 1, wykona polecenie w drugiej linijce (wydrukuje wartość i), a po dotarciu do polecenia NEXT powróci do początku, powtarzając proces z wartością i = 1+2 itp. Można powiedzieć, że komputer ‘rozumie’ znaczenie tego

zdania.

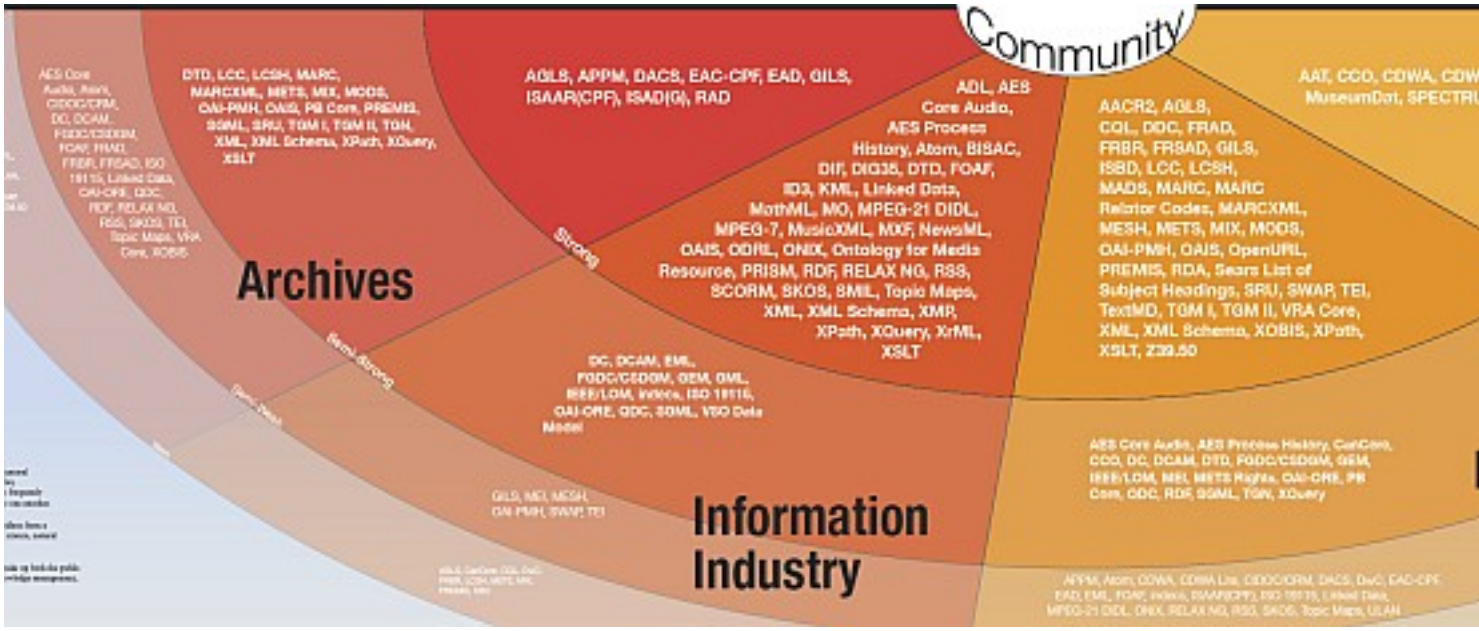
XML

Podstawowym językiem wszystkich standardów metadanych jest XML (eXtensible Markup Language). Język ten posiada bardzo prostą składnię i w zasadzie pozbawiony jest semantyki. Jest on stworzony w ten sposób z założenia po to, aby można było go rozbudowywać ("extensible" w nazwie). Język naturalny nie ma w ogóle dobrego pojęcia na określenie takiego obiektu jak XML, bo przyzwyczailiśmy się rozumieć "język" jako coś, co posiada zarówno składnię jak i semantykę. Pomaga nieco myślenie o XML jako o 'regułach' z których tworzy się dopiero język posiadający znaczenie - czasem używa się również określenia 'format tekstowy'. Prawie wszystkie nowoczesne standardy metadanych mają XML jako podstawę, i tylko uzupełniają o znaczenia formalną strukturę składni.

XML używa znaczników - etykiet - do oznakowania tekstu tak, aby komputer mógł lepiej taki tekst 'zrozumieć' A więc zdanie "Byłem w Warszawie i widziałem Syrenę". można oznakować np tak "Byłem w <miasto>Warszawie</miasto> i widziałem <pomnik>Syrenę</pomnik>". Zdanie "Byłem w Zgierzu Syreną" można oznakować tak: "Byłem w
<miasto>
Zgierzu
</miasto> <samochód>
Syreną
</samochód>
". Składnia XML to etykiety, oznaczane przy pomocy ostrych nawiasów, np.
<miasto>
oznacza "od tego miejsca mówimy o czymś, co ma etykietę
<miasto>
. Wymagana jest etykieta końcowa
</miasto>
która oznacza "w tym miejscu skończył się fragment tekstu oznaczony etykietą
<miasto>
. Pomiędzy etykietami może być dowolny tekst, również dalsze etykiety, co pozwala na budowanie hierarchii.

Drugim pojęciem XML jest atrybut, który można używać do modyfikowania znaczenia etykiety. Np. <miasto rodzaj="stołeczne">Warszawa</miasto> gdzie słowo rodzaj jest atrybutem, a rodz aj="stołeczne"
to przykład składni jego użycia. W tekście zawartym między etykietami może być dowolna ilość (nawet powtarzających się) etykiet. Wewnątrz etykiety może być dowolna ilość, ale nie powtarzających się atrybutów.

To jest w zasadzie wszystko, co trzeba wiedzieć o XML. Sam standard jest nieco bardziej rozbudowany, zawiera dodatkowe reguły, np. o tym jak umieścić znak < między etykietami, czy etykieta może zaczynać się od liczby (nie) itp., ale to są szczegóły które maja znaczenie przy budowaniu XML, ale nie są istotne przy jego czytaniu. Co jest ważniejsze, standard XML nie określa w żaden sposób zawartości etykiet i atrybutów - można tam wstawić wszystko co mieści się w regułach składni. Dlatego tez w XML można wyrazić bardzo dużo - można w nim skatalogować książkę, oznakować jej zawartość, zapisać zawartość bazy danych, i oczywiście zapisać metadane dowolnego obiektu.



Fragment graficznej ilustracji wybranych standardów metadanych. Zawartość: Jenn Riley, projekt: Devin Becker, praca opłacona przez Indiana University Libraries White Professional Development Award.

Składnia i semantyka

Standardy które używają XML to języki które mają i składnię i semantykę. Składnia - to szczegółowe reguły które opisują, jakie elementy i jakie atrybuty muszą być użyte, jakie mogą ale nie muszą być użyte, w jakiej kolejności itp. Składnia może być tylko opisowa (wyrażona w języku naturalnym), ale lepiej jest użyć specjalnego języka do opisu składni, zwanego Schema. Użycie Schema, który sam jest wyrażony w XML, pozwala również komputerom na poznanie reguł składni, i na przykład sprawdzenie czy dany dokument te reguły spełnia. Semantyka

standardu to opis, już zawsze w języku naturalnym, znaczenia i użycia poszczególnych elementów i atrybutów. Dobrze napisana semantyka zawiera nie tylko formalne definicje ale także przykłady ich użycia.

Aby lepiej wyobrazić sobie te różnice, spróbujmy wziąć jako przykład obiekt, książkę p.t. “Bibuła” pisaną przez Józefa Piłsudskiego w 1903 roku. W składni standardu **Dublin Core** (DC) opis książki mógłby wyglądać np. tak:

```
<dc:creator>Józef Piłsudski (1987-1935)<dc:creator>
<dc:title>Bibula</dc:title>
```

(Używamy tu konwencji poprzedzania etykiety skrótem nazwy standardu. Konwencja ta zrozumiała jest też przez komputer, który może w razie potrzeby sięgnąć do formalnego opisu składni danego standardu.)

W składni standardu **MARC** (a ściślej jego wyrażenia w XML, MARCXML) ten sam fragment wygląda tak:

```
<marc:datafield tag="100" ind1="1" ind2=" ">

<marc:subfield code="a">Piłsudski, Józef</marc:subfield>
 <marc:subfield code="d">(1867-1935).</marc:subfield>

</marc:datafield>

<marc:datafield tag="245" ind1="1" ind2="0">

<marc:subfield code="a">Bibuła /</marc:subfield>
```

<marc:subfield code="c">Józef Piłsudski.</marc:subfield>

</marc:datafield>

Składnia jest bardziej złożona, ale możemy rozpoznać te same elementy danych - autora, tytuł, itp.

Podobnie możemy porównać semantykę standardów. Biorąc na przykład etykietę która występuje w wielu standardach, <abstract>, możemy zaleźć definicje dla przykładowych standardów:

- standard Dublin Core: <abstract> - “streszczenie zasobu”.
- standard EAD: <abstract> - “bardzo krótkie streszczenie opisywanych materiałów, używane głównie do zapisu fragmentów informacji bibliograficznych lub historycznych o twórcy i skróconych informacji o zakresie, zawartości, ułożeniu i innych opisowych informacji o jednostce archiwalnej lub jej części”.
- standard MODS: <abstract> - “streszczenie zawartości zasobu”.

Standardy metadanych można pogrupować według różnych kryteriów (wymiarów), np. ich funkcji, domeny, czy społeczności która je używa. Można je także uporządkować według tego, czy zawierają głownie elementy składni, semantyki czy też obu. Na przykład standard AACR2 (Anglo-American Cataloging Rules) jest czysto opisowy - prawie wyłącznie semantyka. Dublin Core jest w większości semantyczny, ale posiada słownik etykiet w XML, a więc pewne minimum składni. Standard EAD, używany przez archiwistów, posiada złożona składnię a także część opisową, a więc jest w miarę kompletnym językiem.

W następnych częściach będziemy pisać o różnych aspektach, wymiarach i funkcjach standardów, szczególnie tych, które są ważne dla archiwów i bibliotek. Zapraszamy do zaglądania do tego blogu.

Marek Zieliński

Artykuł ukazał się 26 marca 2013 w Blogu archiwistów i bibliotekarzy Instytutu Piłsudskiego