Formaty danych strukturalnych w zasobach World Wide Web

Marcin Roszkowski
Katedra Informatologii, Wydział Dziennikarstwa, Informacji i Bibliologii, Uniwersytet Warszawski

Abstrakt

Cel/Teza: Celem artykułu jest charakterystyka i porównanie formatów danych wykorzystywanych do strukturyzacji metadanych dotyczących treści zasobów World Wide Web w ramach systemu języka znacznikowego HTML. Przedmiotem badań są wybrane formaty danych strukturalnych (mikroformaty, RDFa, mikrodane i JSON-LD) oraz zakres ich wykorzystania na potrzeby reprezentacji informacji w środowisku WWW. Punktem wyjścia do rozważań są tezy, że rozdzielenie warstwy treści zasobów sieciowych od sposobu ich prezentacji jest jedną z fundamentalnych cech środowiska World Wide Web oraz że poziom strukturyzacji treści zasobów sieciowych jest czynnikiem determinującym zakres ich wyszukiwalności.

Koncepcja/Metody badań: Scharakteryzowano koncepcję formatów danych strukturalnych (FDS), która zakłada formalną reprezentację informacji o treści zasobów sieciowych bezpośrednio wewnątrz dokumentów HTML. Analizie poddano formaty danych najszerszej wykorzystane w środowisku WWW. Przyjęto również założenie, że FDS może być interpretowany jako środek ekspresji metadanych dokumentu oraz że reprezentacja informacji ma miejsce nie na poziomie samego dokumentu, ale na poziomie faktów go konstytuujących. Analiza FDS została przeprowadzona z punktu widzenia sposobów formalnej ekspresji metadanych (poziom syntaktyczny) oraz zastosowanych do opisu struktur pojęciowych wraz z ich językowymi wykładnikami (poziom semantyczny).

Wyniki i wnioski: FDS przeznaczone do reprezentacji treści stron internetowych zapewniają nowe możliwości strukturyzacji treści w ramach systemu znaczników języka HTML i tym samym rozszerzają zakres możliwych funkcjonalności mechanizmów wyszukiwawczych. Wyniki badań nad implementacją FDS w latach 2012–2016 pokazują wzrost zainteresowania tą formą strukturyzacji treści w zasobach WWW. Największą szczegółowość w reprezentacji wiedzy zapewnia RDFa, jednak to mikrodane wydają się być kompromisem między pożądaną siłą ekspresji a prostotą implementacji, na co wskazują wyniki badań szczegółowych uzyskanych w projekcie Web Data Commons.

Oryginalność/Wartość poznawcza: Przedstawione porównanie wybranych elementów FDS daje obraz ich możliwości w głębokiej strukturyzacji treści zasobów WWW, ze szczególnym uwzględnieniem wykorzystania istniejących schematów metadanych i ontologii. Analiza dokumentacji projektu Web Data Commons z lat 2014–2016 pozwala sądzić, że to mikrodane będą FDS, który będzie miał istotne znaczenie w kontekście zastosowania technologii semantycznych na potrzeby strukturyzacji treści zasobów WWW.

Słowa kluczowe

JSON-LD. Linked Data. Metadane. Mikrodane. Mikroformaty. RDFa. Schema.org. Sieć Semantyczna.

Otrzymany: 30 sierpnia 2017. Zrecenzowany: 16 listopada 2017. Zaakceptowany: 18 grudnia 2017.

1. Wprowadzenie

Jedną z fundamentalnych cech World Wide Web (WWW) jest rozdzielenie warstwy prezentacji informacji od sposobu formatowania i zapisu danych ją tworzących (Guha et al., 2015). Na płaszczyźnie prezentacji informacji mamy do czynienia z interpretacją przez przeglądarki internetowe kodu stron internetowych, zapisanego najczęściej w języku HTML oraz wyświetleniem użytkownikowi treści stron WWW wraz dodatkowymi funkcjonalnościami i elementami interaktywnymi. Poziom formatowania i strukturyzacji informacji to wykorzystanie języków znacznikowych (ang. *markup language*) do organizacji stron internetowych (język HTML – HyperText Markup Language) oraz projektowania ich typografii (np. język CSS – Cascading Style Sheets). HTML jest językiem, który w założeniu ma jednak opisywać strukturę dokumentu, nie zaś jego treść. System jego znaczników odpowiada w większości za kwestie związane z formatowaniem dokumentu i sposobem prezentacji jego treści użytkownikowi za pomocą przeglądarki internetowej. Z punktu widzenia formatowania danych tworzących zawartość stron internetowych, ograniczenia języka HTML powodują, że treść tych dokumentów na poziomie semantycznym nie może być formalnie odwzorowana w celu jej automatycznego przetwarzania. Możliwości języka HTML w kontekście reprezentacji informacji o zawartości stron internetowych kończą się na metadanych dotyczących całego dokumentu, które zapisywane są w jego części nagłówkowej (sekcja <head>). Z punktu widzenia architektury sieciowych systemów informacyjnych, stosowane w nich rozwiązania bazodanowe oferują duże możliwości strukturyzacji przechowywanych danych, lecz w efekcie są one publikowane w postaci stron internetowych za pośrednictwem systemu języka HTML, gubiąc jednocześnie ich pierwotne uporządkowanie. Z punktu widzenia mechanizmów wyszukiwarek internetowych zaindeksowanie dokumentu wiąże się z automatyczną interpretacją ciągu znaków, które tworzą jego treść. Brak informacji o treści dokumentu wyrażonej w formie przetwarzalnej przez aplikację powoduje, że adekwatność indeksowania rzeczowego zależy od efektywności zastosowanych algorytmów przetwarzania języka naturalnego. Taka sama sytuacja ma miejsce w odniesieniu do narzędzi agregujących dane z WWW (takich, jak np. serwisy porównywania cen produktów). Jeżeli zastosowany w danym serwisie internetowym system zarządzania treścią (CMS – Content Management System) nie oferuje dostępu do kolekcji za pośrednictwem interfejsu programistycznego (API – Application Programming Interface), to pozyskanie wybranych treści wiąże się z koniecznością opracowania algorytmu i mechanizmu ekstrakcji danych przeznaczonych dla danego źródła internetowego.

Można więc przyjąć, że poziom strukturyzacji treści zasobów sieciowych jest czynnikiem determinującym zakres funkcjonalności wyszukiwarek internetowych, które je indeksują (Bizer et al., 2012, 3). Z drugiej strony, dążenie do zwiększenia możliwości publikowania informacji na głębokim poziomie strukturyzacji było jednym z początkowych założeń koncepcji WWW (Bizer et al., 2009, 15). Tim Berners-Lee już w 1994 r. założył, że rozwój WWW powinien uwzględnić zmianę sposobu publikowania dokumentów w sieci. Miała ona polegać na formatowaniu stron internetowych w sposób adekwatny do ich percepcji przez użytkownika oraz na zastosowaniu przetwarzalnych maszynowo informacji o charakterze semantycznym (Berners-Lee, 1994, 797). Wizja T. Bernersa-Lee doprowadziła do rozwoju koncepcji Sieci Semantycznej (ang. *Semantic Web*), w której zakładał nałożenie dodatkowej warstwy na istniejące kolekcje sieciowe, zapewniającej formalną reprezentację

informacji zapisanych w ich zasobach. Idea Sieci Semantycznej opiera się na publikowaniu w środowisku sieciowym wzajemnie powiązanych, ustrukturyzowanych zbiorów danych z wykorzystaniem przyjętych za standardowe technologii informacyjnych. Według tej koncepcji (Berners-Lee et al., 2001) w rozproszonych zasobach sieciowych reprezentacja informacji powinna mieć miejsce nie na poziomie dokumentu, lecz na poziomie faktów ekstrahowanych z jego treści. Takie podejście do tworzenia i udostępniania metadanych pozwoli na utworzenie sieci danych (ang. *Web of Data*), która bazować ma na istniejącej sieci dokumentów. Dla WWW standardem opisu zasobów sieciowych nadal pozostaje HTML, natomiast *lingua franca* Sieci Semantycznej jest język Resource Description Framework (RDF). Zaznaczyć należy, że RDF jest językiem reprezentacji wiedzy przeznaczonym do opisu zbiorów danych, nie zaś – tak jak HTML – językiem prezentacji treści stron internetowych.

Jedną z form realizacji Sieci Semantycznej jest metodyka publikowania danych o nazwie Linked Data. Zakłada ona udostępnianie całych zbiorów danych, zgodnie z przyjętymi standardami sieciowymi oraz konieczność ustalania relacji między elementami opisywanymi w tych zbiorach a zewnętrznymi źródłami informacji. Utworzona w ten sposób sieć danych nazywana jest chmurą danych powiązanych (ang. *Linked Data Cloud*). Chociaż jest to wiele obiecująca realizacja wizji Tima Bernersa-Lee, to problem polega na tym, że dane źródłowe muszą być konwertowane zgodnie z założeniami Linked Data oraz udostępniane w pierwszej kolejności dla aplikacji, które będą je przetwarzać. Mamy więc do czynienia z funkcjonowaniem kilku wersji tego samego źródła informacji, np. Wikipedia i jej „semantyczna wersja" DBpedia[1]. Istotne więc wydaje się być pytanie o status istniejących zasobów WWW w kontekście rozwoju technologii semantycznych.

Jednym z rozwiązań problemu ograniczeń HTML w zakresie formalnej strukturyzacji zawartości stron internetowych jest stosowanie formatów danych operujących tzw. semantycznymi znacznikami (ang. *semantic markup*), które stanowią rozszerzenia systemu znaczników tego języka. W odniesieniu do systemu języka znacznikowego HTML termin *znacznik semantyczny* jest rozumiany jako formalny wykładnik reprezentujący element treści dokumentu oraz jako składnik specyfikacji systemu znaczników, którego celem jest semantyczna strukturyzacja zawartości stron WWW.

2. Cel i metodologia badań

Celem artykułu jest charakterystyka i porównanie formatów danych, wykorzystywanych do strukturyzacji treści stron internetowych w ramach systemu języka znacznikowego HTML. Przedmiotem badań są cztery formaty danych strukturalnych (FDS; ang. *structured data formats*), aktualnie wykorzystywane do reprezentacji informacji na stronach internetowych WWW:

(1) mikroformaty;
(2) Recource Description Framework in Attributes (RDFa);
(3) mikrodane;
(4) Java Script Object Notation for Linked Data (JSON-LD).

[1] http://dbpedia.org/

Wybór formatów do analizy oparto na kryterium zakresu ich rozpowszechnienia w środowisku WWW. Podstawę do identyfikacji danych do analiz stanowiły wyniki badań projektu Web Data Commons[2], którego celem była ekstrakcja danych ustrukturyzowanych z korpusu stron internetowych oraz analiza zakresu ich implementacji. W artykule przedstawiono również analizę wyników badań tego projektu w kontekście rozpowszechnienia wspomnianych czterech formatów danych strukturalnych.

Przyjęta metodologia badań zakłada interpretację czterech wspomnianych specyfikacji jako formatów danych, służących do tworzenia metadanych na temat treści stron internetowych i zagnieżdżania tych informacji w strukturze dokumentu HTML. Przeprowadzone analizy uwzględniały aspekt semantyczny i syntaktyczny. W pierwszym przypadku oznacza to dociekania związane z zakresem ekspresywności w danym systemie znaczników semantycznych. W drugim ujęciu, każdy z formatów danych został przedstawiony z punktu widzenia formalnych rozwiązań prezentacji metadanych zagnieżdżanych w strukturze dokumentu HTML. Dla każdej z omawianych specyfikacji wskazano zalety i wady związane z jej implementacją.

Do realizacji celu głównego opracowano zestaw szczegółowych pytań badawczych:

(1) Jaki jest kontekst powstania danego formatu danych strukturalnych?
(2) Jaki model danych reprezentuje dany FDS?
(3) Czy dany FDS wprowadza nowe elementy do systemu języka HTML?
(4) W jaki sposób konstruowane są wykładniki dla elementów metadanych dla każdego z FDS?
(5) Czy dany FDS pozwala na wykorzystanie istniejących schematów metadanych?
(6) Jaki jest szacunkowy zakres wykorzystania danego FDS do reprezentacji informacji na stronach internetowych?

3. Analiza formatów danych strukturalnych

Z punktu widzenia ewolucji koncepcji formatów danych strukturalnych, do najwcześniejszych prób opracowania specyfikacji umożliwiających zagnieżdżanie metadanych w strukturze dokumentu HTML, należą mikroformaty oraz RDFa. Rozwój tych formatów postępował równolegle do ewolucji języka HTML, który stanowić miał ich naturalne środowisko funkcjonowania, ale był zainicjowany przez mniej (mikroformaty) lub bardziej (RDFa) zorganizowane społeczności internautów. Specyfikacja mikrodanych została opracowana w 2009 r. jako odpowiedź twórców języka HTML na zainteresowanie strukturyzacją treści stron internetowych, zaś JSON-LD jest jedną z ostatnich prób standaryzacji reprezentacji informacji w ramach struktury dokumentów HTML. Przyjęta kolejność analiz FDS odpowiada zatem chronologii, w jakiej formaty te opracowano.

3.1. Mikroformaty

Mikroformaty to jedna z pierwszych prób wprowadzenia danych strukturalnych do systemu języka HTML (Bizer et al., 2012, 3; Ronallo, 2012). Powstały one dzięki społecznej

[2] http://www.webdatacommons.org/structureddata/

inicjatywie, której celem było opracowanie szeregu specyfikacji znaczników semantycznych, które w założeniu w prosty sposób mogłyby zostać wykorzystane jako metadane opisujące wybrane fakty publikowane na stronach internetowych.

Celem prac nad mikroformatami było pokonanie ograniczeń języka HTML w zakresie odwzorowania treści stron internetowych, co miało przełożyć się na większą efektywność wyszukiwania tak opisanych zasobów sieciowych. Tłem dla tej inicjatywy były również prace nad realizacją koncepcji Sieci Semantycznej, które skierowane były na popularyzację języka XML w środowisku sieciowym, opublikowanie standardu reprezentacji wiedzy RDF oraz ewolucja samego języka HTML. W 2000 r. opublikowano specyfikację XHTML (Extensible Hypertext Markup Language), która nie była kolejną wersją języka HTML, lecz jego specyfikacją, zgodnie z zaleceniami XML. HTML wywodzi się bezpośrednio z języka Standard Generalized Markup Language (SGML), który reprezentuje inny model niż XML, na którym to oparto XHTML. Prace nad tym standardem prowadzono przez następne kilka lat, lecz XHTML spotkał się z krytyką twórców aplikacji sieciowych i ostatecznie porzucono ten projekt na rzecz piątej wersji języka HTML (HTML5), którą udostępniono w 2014 r. Z punktu widzenia ewolucji technologii sieciowych inicjatywa mikroformatów powstała w momencie, w którym z jednej strony, HTML radykalnie ograniczał możliwość strukturyzacji danych w obrębie strony internetowej, a z drugiej, nowe rozwiązania (XML, RDF) były na tyle skomplikowane i kosztowne, że ich potencjalne wdrożenie na potrzeby projektowania serwisów internetowych nie wchodziło wówczas w grę. Mikroformaty były więc odpowiedzią na realizację wizji Sieci Semantycznej, ale w warstwie powierzchniowej, tzn. na poziomie struktury dokumentów HTML, nie zaś baz danych.

Geneza mikroformatów sięga 2003 r., kiedy to opracowano pierwszą z jego specyfikacji – XHTML Friends Network (XFN). Był to wykaz znaczników, który służył do ekspresji relacji interpersonalnych w odniesieniu do autorów i czytelników intensywnie wówczas rozwijającej się blogosfery. XFN opracowany przez Global Multimedia Protocols Group (2003) pozwalał na specyfikację hiperłączy między blogami lub osobistymi stronami internetowymi poprzez możliwość wskazania rodzaju relacji zachodzących między ich autorami (np. relacje przyjacielskie, rodzinne, zawodowe). Dla tworzenia mikroformatów charakterystyczne było oddolne, społeczne podejście do rozwoju tego rodzaju specyfikacji w przeciwieństwie do instytucjonalnej standaryzacji schematów metadanych. Platformą wymiany informacji i dokumentacji specyfikacji mikroformatów stał się utworzony w 2005 r. serwis internetowy microformats.org. Tak więc cały proces prac nad specyfikacją i jej publikowaniem odbywał się zdalnie. Za motto tej inicjatywy przyjęto stwierdzenie: „zaprojektowane w pierwszej kolejności dla ludzi, w drugiej dla maszyn" (ang. *designed for humans first and machines second*). Konsekwencją takiego podejścia jest pragmatyzm, który jest widoczny zarówno na płaszczyźnie specyfikacji tych formatów danych, jak i w sposobie realizacji prac nad nimi.

3.1.1 Poziom semantyczny

Semantyka mikroformatów opiera się na gotowych zestawach znaczników identyfikujących klasy i ich własności, zaprojektowanych i zaakceptowanych przez społeczność zgromadzoną wokół tej inicjatywy. Mamy więc do czynienia ze specyfikacjami metadanych dla różnych kategorii pojęciowych występujących w treści stron internetowych, które są przeznaczone do zagnieżdżania w strukturze dokumentu HTML. Z punktu widzenia organizacji struktur pojęciowych, dla których formalne wykładniki zapewnia dana specyfikacja mikroformatu,

mamy do czynienia z układem hierarchicznym (Sporny, 2015), którego trzonem jest rodzaj opisywanego elementu (np. osoba, wydarzenie, produkt, recenzja), a znaczniki z nim związane reprezentują jego atrybuty i stanowią jego semantyczne otoczenie. Tabela 1 zawiera przykładowe specyfikacje mikroformatów wraz ze wskazaniem zakresu ich stosowania (zob. też Tab. 6).

Tab. 1. Przykładowe specyfikacje mikroformatów

Mikroformat	**Zakres stosowania**
hCard[1]	metadane dla osób, organizacji
hCalendar[2]	metadane dla wydarzeń
hReview[3]	metadane dla recenzji (produktów i usług)
hAtom[4]	metadane dla wpisów na blogach internetowych
hProduct[5]	metadane dla produktu handlowego
hReview[6]	metadane dla opinii, recenzji produktu lub usługi

[1] http://microformats.org/wiki/hCard
[2] http://microformats.org/wiki/hCalendar
[3] http://microformats.org/wiki/hReview
[4] http://microformats.org/wiki/hAtom
[5] http://microformats.org/wiki/hproduct
[6] http://microformats.org/wiki/hreview

Specyfikacja mikroformatów zakłada funkcjonowanie atrybutów prostych i ustrukturyzowanych (Tomberg & Laanpere, 2009). W pierwszym przypadku są to elementy metadanych. W drugim – mają one postać elementów, które reprezentują dany atrybut, ale jego zakres znaczeniowy można dodatkowo uszczegóławiać. Na przykład, mikroformat hCard, przeznaczony do zapisu ustrukturyzowanych informacji na temat osób i instytucji, zakłada funkcjonowanie atrybutów prostych, m.in.:

- fn: nazwa prosta
- email: adres e-mail
- tel: numer telefonu
- url: adres URL strony domowej
- bday: data urodzin

ale również elementy ustrukturyzowane, m.in.:

- adr: adres zamieszkania
  - street-address: ulica
  - locality: miejscowość
  - region: region, stan
  - postal-code: kod pocztowy
  - country-name: państwo
- n: nazwa (element ustrukturyzowany)
  - honorific-prefix: zwroty grzecznościowe (np. Pan, Pani)
  - given-name: imię
  - additional-name: drugie imię
  - family-name: nazwisko
  - honorific-suffix: tytuły (np. Dr, Prof.).

Koncepcja atrybutów ustrukturyzowanych nawiązuje do kwalifikatorów w schemacie metadanych Dublin Core[3], które były przez krótki czas stosowane przed wprowadzeniem jego rozszerzonej wersji – Terminów Metadanych DCMI[4]. W modelu mikroformatów stosuje się koncepcję obligatoryjności i fakultatywności elementów metadanych. Specyfikacja danego mikroformatu w sposób jawny określa, które z elementów metadanych mają status obowiązkowy.

3.1.2 Poziom syntaktyczny

Na poziomie syntaktycznym w mikroformatach została wykorzystana podstawowa składnia HTML. Oznacza to, że nie wprowadzono w nich nowych elementów do systemu znaczników języka HTML. Specyfikacja mikroformatu zakłada funkcjonowanie klas, czyli rodzajów opisywanych obiektów, ich własności oraz zestawu reguł związanych z formalną ekspresją metadanych. Mikroformaty wykorzystują selektory języka HTML, najczęściej znaczniki reprezentujące elementy strukturalne dokumentu (np. *h1* – nagłówek pierwszego stopnia, *p* – paragraf, *a* – hiperłącze lub elementy grupujące – np. *div*), których znaczenie jest definiowane najczęściej za pomocą atrybutu znacznika class z wykorzystaniem słownictwa danego mikroformatu. Znacznik *class* w języku HTML jest przewidziany przede wszystkim definiowania elementów w strukturze dokumentu HTML, dla których określono specyficzny sposób ich graficznej prezentacji z wykorzystaniem Kaskadowych Arkuszy Stylów (CSS). W przypadku mikroformatów został on zaadaptowany do deklaracji rekordu danych strukturalnych. Niżej przedstawiono przykładowe formatowanie akapitu tekstu zawierającego dane na temat Jana Kowalskiego oraz jego afiliacji za pomocą mikroformatu hCard.

Wersja wyświetlana użytkownikowi:

Jan Kowalski
Uniwersytet Warszawski
jan@kowalski.pl
Krakowskie Przedmieście 26/28
Warszawa , 00–927 Polska

Wersja sformatowana

```
<div id="hcard-Jan-Kowalski" class="vcard">
<span class="fn">Jan Kowalski</span>
<div class="org">Uniwersytet Warszawski</div>
    <a class="email">jan@kowalski.pl</a>
<div class="adr">
    <div class="street-address">Krakowskie Przedmieście 26/28</div>
            <span class="locality">Warszawa</span>
            <span class="postal-code">00–927</span>
            <span class="country-name">Polska</span>
</div>
```

Rekord metadanych na temat Jana Kowalskiego sformatowano bezpośrednio w treści dokumentu za pomocą znacznika grupującego div, gdzie nadano mu identyfikator *id="hcard-Jan-Kowalski"*. W dalszej części rekordu widoczne są wykładniki dla atrybutów prostych (np. *fn*, org) oraz ustrukturyzowanych (*adr*).

[3] http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/
[4] http://www.bn.org.pl/download/document/1253606451.pdf

Zastosowanie mikroformatów uwzględnia również odwzorowanie dodatkowych informacji za pomocą znaczników HTML *rel* oraz *href*, które umożliwiają określenie relacji między elementem treści danego dokumentu a zewnętrznym zasobem sieciowym. Na przykład, dodając zapis w treści dokumentu informujący, że został on opublikowany na licencji „Uznanie autorstwa 2.0 Ogólny (CC BY 2.0)", można formalnie odwzorować ten fakt za pomocą deklaracji z wykorzystaniem mikroformatu „license", tworząc dodatkowo hiperłącze do treści licencji:

```
<a href="http://creativecommons.org/licenses/by/2.0/" rel="license">Opublikowano na licencji:
Uznanie autorstwa 2.0 Ogólny (CC BY 2.0)</a>
```

3.1.3 Zalety i ograniczenia

Do zalet mikroformatów zalicza się ich prostotę, zarówno na płaszczyźnie interpretacji specyfikacji, jak również ich wdrażania w serwisach WWW, zgodność z funkcjonującymi standardami sieciowymi (mikroformaty wykorzystują system znaczników języka HTML), a także duże wsparcie społeczności. Krytyka mikroformatów dotyczy przede wszystkim ich niedostatecznej siły semantycznej oraz **braku formalnego modelu** reprezentacji informacji (Tomberg & Laanpere, 2009, 104). W pierwszym przypadku oznacza to małą liczbę specyfikacji i tym samym poważne ograniczenia w szczegółowości odwzorowania treści zasobów WWW. Brak formalnego modelu danych, który leżałby u podstaw mikroformatów, powoduje problemy związane z zastosowaniem automatycznego przetwarzania i analizy spójności danych.

Z punktu widzenia sposobu identyfikacji elementów i atrybutów istotne jest odnotowanie faktu, że żaden z elementów rekordu nie jest identyfikowany za pomocą standardu URI. Oznacza to, że opisywany obiekt, jego kategoryzacja oraz wszystkie elementy metadanych mają postać danych tekstowych. W konsekwencji opisywany obiekt nie ma tożsamości sieciowej (nie posiada identyfikatora sieciowego) i jest rozpoznawalny wyłącznie w kontekście dokumentu HTML. Model mikroformatów ma ograniczone możliwości związane z jawnym określeniem zasad kodowania danych tekstowych (ang. *typed literals*; np. jednostki miar i czasu) i pozwala jedynie na specyfikację języka tekstu elementu metadanych.

3.2. RDFa

Specyfikacja Resource Description Framework in Attributes (RDFa) została opracowana w 2004 r., a cztery lata później została opublikowana jako oficjalna rekomendacja konsorcjum W3C (Bizer et al., 2012, 8). W założeniu RDFa miał być formatem wykorzystywanym wyłącznie w ramach języka znacznikowego XHTML, ale wraz z ewolucją HTML format ten został dostosowany do aktualnie funkcjonującej wersji HTML5 (Herman et al., 2015). RDFa opracowano jako serializację[5] standardu reprezentacji wiedzy RDF, przeznaczoną do zagnieżdżania ustrukturyzowanych danych wewnątrz dokumentów HTML (Bizer et al., 2012, 8). RDF jest rekomendowany przez konsorcjum World Wide Web jako standard reprezentacji wiedzy w Sieci Semantycznej oparty na modelu grafowym. Wprowadzono

[5] Pojęcie serializacji oznacza tutaj jawny i udokumentowany formalny sposób ekspresji. Np. format MARC 21 posiada serializację w postaci specyfikacji MARC/XML, ale również formatu wymiennego ISO 2709.

w nim koncepcję tzw. trójki RDF (ang. *RDF Triple*) jako elementarnej jednostki wypowiedzi, w której wyróżnia się *przedmiot* opisu (ang. *subject*), *predykat* (ang. *predicate*; atrybut lub wykładnik relacji semantycznej) oraz *obiekt* (ang. *object*; będący wartością opisywanej cechy lub przedmiotem powiązanym relacją reprezentowaną przez predykat). RDFa jako serializacja RDF na poziomie konceptualnym realizuje założenia RDF i pozwala na zagnieżdżanie metadanych w postaci trójek RDF wewnątrz systemu znacznikowego języka HTML. RDF jest przeznaczony do formalnej reprezentacji danych na potrzeby baz wiedzy w ramach Sieci Semantycznej, natomiast celem RDFa jest wykorzystanie ekspresywności tego modelu na potrzeby opisu zwartości stron WWW.

3.2.1 Poziom semantyczny

Podstawową cechą RDFa w kontekście semantycznej warstwy reprezentacji informacji jest brak z góry narzuconego zestawu elementów metadanych. Oznacza to możliwość wykorzystania dowolnego schematu metadanych lub ontologii sieciowej, które posiadają swoją formalną specyfikację. Warunek formalnej reprezentacji schematu metadanych, który ma być wykorzystany w opisie za pomocą RDFa oznacza dostęp do niego za pośrednictwem jego unikalnej sieciowej przestrzeni nazw (ang. *namespace*) i tym samym stosowanie standardu URI do identyfikacji jego elementów strukturalnych[6]. Jest to niewątpliwa zaleta RDFa, ponieważ pozwala na ponowne użycie (ang. *re-use*) istniejących specyfikacji metadanych i brak konieczności opracowania nowych, wyłącznie na potrzeby tego formatu danych strukturalnych. Dodatkowo w ramach jednego rekordu metadanych (zestawu trójek RDF) istnieje możliwość jednoczesnego stosowania wielu takich specyfikacji.

3.2.2 Poziom syntaktyczny

Na poziomie syntaktycznym w RDFa wprowadzono szereg nowych atrybutów do składni HTML, których celem jest dostosowanie modelu RDF do koncepcji zagnieżdżania metadanych w strukturze HTML. RDFa funkcjonuje w dwóch wersjach – wersji podstawowej RDFa Core 1.1[7] oraz wersji uproszczonej – RDFa Lite[8]. Różnica polega na liczbie elementów syntaktycznych, a więc tym samym na poziomie szczegółowości opisów. RDFa Lite wprowadzono w celu ułatwienia twórcom serwisów internetowych implementacji tego FDS, a kryteria doboru elementów miały charakter pragmatyczny. Dokumentacja RDFa Lite zawiera stwierdzenie, że ta specyfikacja będzie adekwatna w 80% przypadków związanych z opracowaniem semantycznych znaczników dla zawartości stron WWW. W celu ilustracji rozwiązań syntaktycznych RDFa w dalszej części tej sekcji zostaną one przedstawione na podstawie jego uproszczonej wersji. W RDFa Lite wprowadzono pięć nowych atrybutów o charakterze syntaktycznym[9]:

(1) *vocab* – element systemu znacznikowego służący do identyfikacji bazowej przestrzeni nazw (URI) schematu metadanych, z którego elementy będą wykorzystywane w opisie;

[6] Np. schemat metadanych Dublin Core Metadata Element Set jest zarejestrowany w przestrzeni nazw – http://purl.org/dc/elements/1.1/, w której każdy ze zidentyfikowanych w nim atrybutów posiada swój unikalny identyfikator URI (np. tytuł – http://purl.org/dc/elements/1.1/title).

[7] https://www.w3.org/TR/rdfa-core/

[8] https://www.w3.org/TR/rdfa-lite/

[9] RDFa w wersji podstawowej wprowadza osiem nowych elementów syntaktycznych do składnik HTML.

(2) *prefix* – element systemu znacznikowego służący do wprowadzenia akronimu dla dodatkowych przestrzeni nazw stosowanych w opisie w celu skrócenia i uproszczenia sposobu odwoływania się do elementów metadanych;
(3) *typeof* – element systemu znacznikowego służący do kategoryzacji opisywanego obiektu (np. osoba, wydarzenie). Wartość tego elementu pochodzi z wcześniej zadeklarowanej (element *vocab*) przestrzeni nazw dla schematu metadanych;
(4) *property* – element systemu znacznikowego służący do identyfikacji odwzorowywanej własności obiektu. Nazwa tego elementu pochodzi z wcześniej zadeklarowanej (element *vocab*) przestrzeni nazw dla schematu metadanych, a jego wartość w zależności od typu atrybutu przyjmuje postać ciągu znaków lub URI;
(5) *resource* – element systemu znacznikowego służący do identyfikacji opisywanego obiektu.

Poniżej przedstawiono formatowanie akapitu tekstu zawierającego wcześniej przywoływane dane teleadresowe na temat Jana Kowalskiego za pomocą formatu danych RDFa:

```
<div vocab="http://schema.org/" resource="Jan-Kowalski" typeof="Person">
<span property="name">Jan Kowalski</span>
<span property="email">jan@kowalski.pl</span>
<div property="affiliation" resource="UW" typeof="Organization">
    <span property="name">Uniwersytet Warszawski</span>
<div property="address" resource="UW-adres" typeof="PostalAddress">
<span property="streetAddress">Krakowskie Przedmieście 26/28</span>
                <span property="addressLocality">Warszawa</span>
                <span property="postalCode">00–927</span>
                <span property="addressCountry">Polska</span></div>
</div>
```

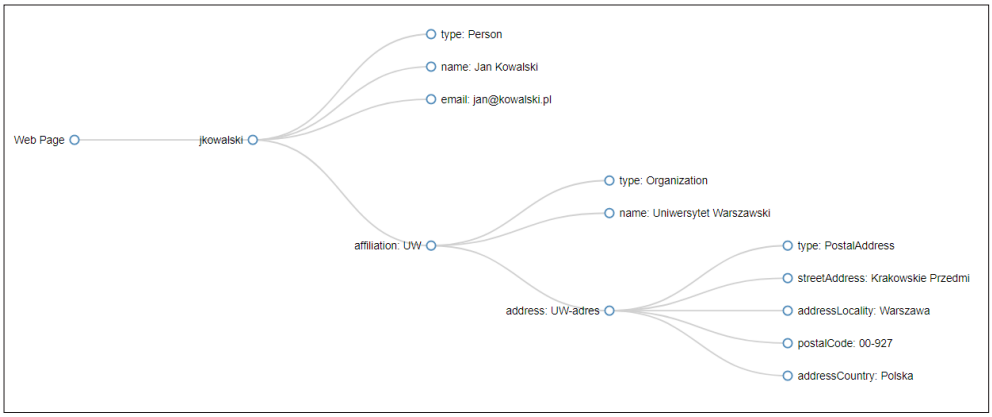
Przedstawiony powyżej zestaw metadanych w formacie RDFa rozpoczyna się od deklaracji wykorzystanego w dalszej części zewnętrznego schematu metadanych (*vocab*), którym w tym przykładzie jest specyfikacja schema.org[10], określenia przedmiotu opisu (*resource*) oraz wskazania jego wewnętrznego identyfikatora dla osoby Jana Kowalskiego (*resource="Jan-Kowalski"*), a także formalnej kategoryzacji opisywanego obiektu, tzn. stwierdzenia, że jest on instancją, tj. wystąpieniem klasy człowiek (*typeof="Person"*). W dalszej części każda wartość atrybutu jest poprzedzona identyfikatorem *property* oraz wykładnikiem atrybutu ze schematu schema.org. Przedmiotem opisu jest zadeklarowany w pierwszej linii obiekt *Jan-Kowalski*, ale zestaw metadanych jest dodatkowo ustrukturyzowany i zawiera również informacje o obiekcie *Uniwersytet Warszawski*, który reprezentuje afiliację *Jana Kowalskiego*. Obiekt ten zidentyfikowano (*resource*) i skategoryzowano jako instancję klasy organizacja (*typeof="Organization"*). Na jego opis składają się elementy proste (nazwa – *name*) oraz ustrukturyzowane (*PostalAddress*) – szczegółową specyfikację danych adresowych przedstawia rysunek 1.

RDFa, tak jak i RDF, wyznacza pewien formalny schemat reprezentacji wiedzy. Jego słownictwo, oprócz podstawowych deklaracji ontologicznych, jest wykorzystywane głównie do organizacji składni. Do niewątpliwych zalet RDFa należy możliwość zastosowania zewnętrznych schematów metadanych i ontologii sieciowych poprzez odwoływanie się do ich zawartości za pośrednictwem URI dla danej przestrzeni nazw. Istnieje więc możliwość

[10] http://schema.org/

jednoczesnego korzystania z wielu takich specyfikacji. Pozwala to na wysoki stopień szczegółowości reprezentacji informacji, ale wiąże się z koniecznością każdorazowej weryfikacji statusu ontologicznego wykorzystanych atrybutów i klas. Poniżej przedstawiono formatowanie tego samego tekstu z zastosowaniem RDFa, gdzie użyto elementy schematu metadanych schema.org oraz ontologii Friend of a Friend (FOAF)[11]:

```
<div prefix="schema: http://schema.org/ foaf: http://xmlns.com/foaf/0.1/" resource="Jan-Kowalski" typeof="foaf:Person">
<span property="foaf:name">Jan Kowalski</span>
<span property="foaf: mbox">jan@kowalski.pl</span>
<div property="schema:affiliation" resource="UW" typeof="schema:Organization">
    <span property="foaf:name">Uniwersytet Warszawski</span>
<div property="schema:address" typeof="schema:PostalAddress" resource="UW-adres>
<span property="schema:streetAddress">Krakowskie Przedmieście 26/28</span>
            <span property="schema:addressLocality">Warszawa</span>
            <span property="schema:postalCode">00–927</span>
            <span property="schema:addressCountry">Polska</span></div>
</div>
```



Rys. 1. Graficzna reprezentacja deklaracji w RDFa

Ontologia FOAF jest jedną z bardziej rozpowszechnionych formalnych struktur informacji osobowej. Jest wykorzystywana do tworzenia tzw. semantycznych wizytówek sieciowych, w których treści można zawrzeć zarówno podstawowe dane osobowe, jak również specyfikować relacje z innymi ludźmi (np. znajomości). W powyższym przykładzie nie zadeklarowano bazowej przestrzeni nazw (*vocab*), lecz wykorzystano wykładnik *prefix* do wskazania na specyfikacje metadanych, które będą użyte do reprezentacji informacji. Imię i nazwisko (*foaf:name*) oraz adres e-mail (*foaf:mbox*) odwzorowano za pomocą wykładników klas (*typeof="foaf:Person"*) i atrybutów z ontologii FOAF. Ontologia ta nie pozwala jednak na dokładne odwzorowanie danych teleadresowych, stąd informacje na temat Uniwersytetu Warszawskiego zapisano za pomocą wykładników atrybutów i klas ze schematu schema.org.

[11] http://xmlns.com/foaf/spec/

3.2.3 Zalety i ograniczenia

Oprócz możliwości łączenia wielu formalnych specyfikacji metadanych i ontologii w ramach zestawu deklaracji, wartą podkreślenia zaletą RDFa jest operowanie identyfikatorami sieciowymi URI ze wskazanych w opisie przestrzeni nazw. A zatem, zastosowany w przykładzie skrócony zapis *foaf:Person* dla kategoryzacji Jana Kowalskiego, formalnie oznacza, że obiekt Jan-Kowalski opisany na danej stronie internetowej jest wystąpieniem klasy osoba w ontologii FOAF, która posiada swoje URI – http://xmlns.com/foaf/spec/#term_Person, a obiekt UW opisany na tej samej stronie WWW jest wystąpieniem klasy organizacja w myśl schematu schema.org, która również posiada swoje unikalne URI – http://schema.org/Organization. Taka formalna kategoryzacja treści zasobów WWW oraz reprezentacja z wykorzystaniem schematów metadanych i ontologii z jednej strony pozwala na precyzyjne indeksowanie zasobów WWW przez wyszukiwarki internetowe, a z drugiej na agregację danych i ich konwersję do postaci baz wiedzy, np. do formatu RDF.

Zastosowanie formatu RDFa może jednak napotkać problemy po stronie pragmatyki. Format ten daje duże możliwości na płaszczyźnie semantycznej, co w warstwie syntaktycznej przejawia się dużą elastycznością. Z drugiej strony, Bethany Wetherill (2014) twierdzi, że wyniki badań wskazują na duży odsetek błędów popełnianych przez twórców serwisów WWW w praktycznym operowaniu wieloma narzędziami reprezentacji informacji oraz odwoływaniu się do nich w deklaracjach RDFa.

3.3. Mikrodane

Mikrodane to format danych strukturalnych, który powstał w 2009 r. w ramach prac nad piątą wersją języka HTML oraz w wyniku krytyki środowiska, dotyczącej włączenia RDFa do specyfikacji HTML (Ronallo, 2012). W założeniu mikrodane miały być alternatywnym sposobem zagnieżdżania semantycznych znaczników w strukturze dokumentów HTML (Bizer et al., 2012, 11), który da więcej możliwości niż mikroformaty, ale będzie prostszy niż RDFa. Format ten, w przeciwieństwie do dwóch omówionych wcześniej, jest elementem języka HTML. Nie stanowi on nadbudowy dla systemu znaczników HTML, tak jak to ma miejsce w przypadku mikroformatów i RDFa, lecz posiada status modułu, który rozszerza jego możliwości zgodnie z założeniami specyfikacji HTML5 o reprezentację danych strukturalnych wewnątrz struktury stron WWW (Sikos, 2015, 35). Format mikrodane powstał w momencie, w którym środowisko twórców serwisów internetowych zainteresowane wprowadzaniem danych strukturalnych miało już kilkuletnie doświadczenie zarówno w adaptacji mikroformatów jak i RDFa. Jest to więc wynik z jednej strony, świadomej ewolucji języka HTML, a z drugiej, refleksji o charakterze pragmatycznym, w której pod uwagę wzięto potencjalne i realne trudności związane z wdrażaniem szczególnie RDFa. I chociaż dla technologii semantycznych format RDF ma kluczowe znaczenie, to jego serializacja w postaci RDFa nie spotkała się z entuzjazmem projektantów serwisów WWW, chociaż z perspektywy kilku lat funkcjonowania mikrodanych, to RDFa uznaje się za standard dojrzalszy (Wetherill, 2014).

3.3.1 Poziom semantyczny

W modelu konceptualnym, leżącym u podstaw mikrodanych, zakłada się funkcjonowanie zbioru par atrybut-wartość nazywanych tutaj obiektami (ang. *item*). Dany obiekt posiada

dodatkowo deklarację ontologiczną (*item type*) wskazującą na kategorię pojęciową, którą reprezentuje opisywany zasób informacyjny (np. człowiek, organizacja, książka) oraz może zawierać tzw. globalny identyfikator (*item id*), który powinien być wyrażony za pomocą standardu URI. Pierwszy element z pary atrybut-wartość określa własność (ang. *property*) opisywanego obiektu. Może ona przyjąć jedną lub wiele wartości. Wartości są odwzorowywane za pomocą ciągu znaków (ang. *string*) lub mają postać złożoną i tworzą je zestawy par atrybut-wartość. Istnieje również możliwość wyrażania wartości danego atrybutu poprzez odwołanie do innego obiektu (McCathie Nevile & Brickley, 2017). Przykładem własności może być oznaczenie odpowiedzialności dla dokumentu, które może mieć tyle wartości, ile było podmiotów zaangażowanych w jego utworzenie. Atrybut ten może przyjąć wartość prostą w postaci ciągu znaków (np. „Jan Kowalski") lub złożoną w postaci odwołania do obiektu o unikalnym identyfikatorze, w którego opisie będzie kolejny zestaw par atrybut-wartość (np. imię: Jan; nazwisko: Kowalski).

Format mikrodane, tak jak RDFa, nie narzuca określonego zestawu metadanych. Podobnie jak w przypadku RDFa, i tutaj można wykorzystać istniejące specyfikacje metadanych i ontologii sieciowych, które posiadają zarejestrowane przestrzenie nazw. Duży udział w powstaniu mikrodanych miała inicjatywa przedstawicieli największych dostawców wyszukiwarek internetowych – Google, Bing, Yahoo i Yandex, w ramach prac nad wspólną specyfikacją metadanych przeznaczoną dla formatów danych strukturalnych. Efektem tej współpracy jest projekt schema.org, którego celem było opracowanie prostej specyfikacji metadanych o szerokim zakresie stosowania do opisu najczęściej występujących typów obiektów w treści stron internetowych (np. osoby, miejsca, wydarzenia, produkty) z wykorzystaniem formatów danych strukturalnych. Schema.org opublikowano w 2011 r. i obecnie liczy on 638 klas i 965 atrybutów i relacji (Guha et al., 2015). Specyfikacja ta została opracowana dla twórców serwisów internetowych jako stosunkowo proste narzędzie do semantycznej strukturyzacji treści stron internetowych w celu optymalizacji ich indeksowania przez wyszukiwarki internetowe i tym samym w celu zwiększenia trafności wyszukiwania informacji. Tym samym, schema.org stał się niejako pierwszym wyborem dla formatu mikrodane na płaszczyźnie semantycznej.

3.3.2 Poziom syntaktyczny

Mikrodane jako immanentny element strukturalny języka znacznikowego HTML5 wprowadza do jego systemu pięć nowych elementów (McCathie Nevile & Brickley, 2017):

(1) *itemscope* – element systemu znacznikowego służący do identyfikacji nowego obiektu, którego opis będzie wyrażony za pomocą par atrybut-wartość;
(2) *itemtype* – element systemu znacznikowego służący do kategoryzacji opisywanego obiektu (np. osoba, wydarzenie);
(3) *itemid* – element systemu znacznikowego służący do identyfikacji opisywanego obiektu z wykorzystaniem identyfikatora URI;
(4) *itemprop* – element systemu znacznikowego służący do identyfikacji odwzorowywanej własności obiektu;
(5) *itemref* – element systemu znacznikowego służący do grupowania par atrybut-wartość w celu organizacji deklaracji wewnątrz opisu.

Najważniejsze elementy mikrodanych przeznaczone do odwzorowania informacji to *itemscope*, *itemtype* oraz *itemprop*. Poniżej przedstawiono formatowanie akapitu tekstu

zawierającego wcześniej przywoływane dane teleadresowe na temat Jana Kowalskiego za pomocą mikrodanych z wykorzystaniem elementów metadanych ze schematu schema.org:

```
<div itemscope itemtype="http://schema.org/Person"
itemid="http://przyklad.pl/JKowalski" >
<span itemprop="name">Jan Kowalski</span>
<span itemprop ="email">jan@kowalski.pl</span>
<div itemprop="affiliation" itemscope itemtype="http://schema.org/Organization" itemid="http://
przyklad.pl/UW" >
    <span itemprop="name">Uniwersytet Warszawski</span>
<div itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
<span itemprop="streetAddress">Krakowskie Przedmieście 26/28</span>
            <span itemprop="addressLocality">Warszawa</span>
                    <span itemprop="postalCode">00–927</span>
            <span itemprop="addressCountry">Polska</span></div>
</div>
```

Przedmiotem opisu w przykładzie jest obiekt (*itemscope*) o identyfikatorze (*itemid*) – http://przyklad.pl/JKowalski. Deklaracja rozpoczynająca opis zawiera również informacje o statusie ontologicznym obiektu poprzez wskazanie, że jest on instancją klasy człowiek (*itemtype="http://schema.org/Person"*) zidentyfikowanej w schemacie schema.org. W dalszej części występują pary atrybut-wartość odwzorowujące imię i nazwisko oraz adres e-mail. Afiliacja Jana Kowalskiego została odwzorowana poprzez ustalenie relacji między obiektem Jan Kowalski (itemid="http://przyklad.pl/JKowalski") a obiektem Uniwersytet Warszawski (itemid="http://przyklad.pl/UW"), której wykładnikiem jest atrybut *affiliation*. Uniwersytet Warszawski zyskał w tym opisie własną tożsamość poprzez nadanie mu (lub odwołanie się do jego) identyfikatora oraz został skategoryzowany jako wystąpienie klasy organizacja (itemtype="http://schema.org/Organization"). Informacje na temat instytucji afiliującej zostały odwzorowane w sposób prosty (nazwa) oraz poprzez wyodrębnienie obiektu złożonego adres, który w ontologii schema.org reprezentuje klasa http://schema.org/PostalAddress. Taki zabieg zastosowano zarówno w przypadku użycia schema.org w RDFa jak i w mikrodanych. Z formalnego punktu widzenia schema.org nie pozwala na użycie atrybutów teleadresowych – nazwa ulicy (*streetAddress*) , lokalizacja (*addressLocality*), kod pocztowy (*postalCode*) oraz państwo (*addressCountry*) bezpośrednio do opisu organizacji, którą w tym przypadku reprezentuje Uniwersytet Warszawski. Są to własności zdefiniowane dla klasy adres pocztowy (*PostalAddress*), która służy do reprezentacji wartości atrybutu adres (*address*). Specyfikacja schema.org pozwala bowiem na dwa sposoby odwzorowania atrybutu adres. Pierwszym z nich jest nieustrukturyzowany ciąg znaków, który zawiera dane teleadresowe, drugim – wersja ustrukturyzowana z wykorzystaniem typu klasy adres pocztowy (*PostalAddress*). Stosując pierwsze rozwiązanie, dane te mogą przyjąć następujące formatowanie:

```
<div itemscope itemtype="http://schema.org/Person"
itemid="http://przyklad.pl/JKowalski" >
<span itemprop="name">Jan Kowalski</span>
<span itemprop ="email">jan@kowalski.pl</span>
<div itemprop="affiliation" itemscope itemtype="http://schema.org/Organization" itemid="http://
przyklad.pl/UW" >
    <span itemprop="name">Uniwersytet Warszawski</span>
<span itemprop="address"> Krakowskie Przedmieście 26/28, 00–927 Warszawa, Polska</span>
</div>
```

3.3.3 Zalety i ograniczenia

Mikrodane jako format danych strukturalnych postrzegane są jako wynik konsensusu między dążeniem do pogłębionej strukturyzacji treści stron internetowych a prostotą systemu znaczników semantycznych. Tomberg i Laanpere (2009) traktują mikrodane jako hybrydę, która powstała w wyniku połączenia rozwiązań zaczerpniętych z mikroformatów, RDFa i HTML5. Nie ulega jednak wątpliwości, że wsparcie mikrodanych przez Google, Bing, Yahoo i Yandex poprzez rekomendację tego FDS dla implementacji schema.org pozwala sądzić, że będzie on coraz szerzej stosowany do strukturyzacji treści stron internetowych. Do zalet mikrodanych należy możliwość stosowania wielu schematów metadanych i ontologii oraz wykorzystanie identyfikatorów sieciowych (URI) dla opisywanych obiektów. Model mikrodanych nawiązuje do modelu RDF, który leży u podstaw RDFa, co traktuje się jako zaletę, z którą wiąże się możliwość stosunkowo łatwego generowania trójek RDF z tak ustrukturyzowanej zawartości stron WWW na potrzeby tworzenia lub wzbogacania semantycznych baz danych (Sporny, 2015; Tomberg & Laanpere, 2009; Wetherill, 2014). Prostota, czyli to, co jedni (np. Guha et al., 2015) postrzegają jako zaletę mikrodanych, przez innych (np. Wetherill, 2014) jest traktowana również jako ograniczenie. Większość krytyki mikrodanych jest skoncentrowana na fakcie, że pragmatyczne pobudki i mocne powiązanie ze schema.org powodują, iż format ten nie nadaje się do opisów złożonych o dużym stopniu analityczności, np. publikacji naukowych czy treści z sektora e-government.

3.4. JSON-LD

Format danych strukturalnych JSON-LD (Java Script Object Notation for Linked Data) powstał w odpowiedzi na zainteresowanie środowiska twórców usług sieciowych (aplikacje sieciowe, serwisy internetowe) implementacją metodyki publikowania danych według założeń Linked Data. Punktem wyjścia było poszukiwanie adekwatnego rozwiązania w postaci rozszerzenia możliwości ekspresji formatu Java Script Object Notation (JSON). Jest on szeroko stosowanym formatem danych, który zapewnia efektywną reprezentację danych przesyłanych z serwera obsługującego dany serwis WWW do przeglądarki internetowej, wyświetlającej jego zawartość i zapewniającej interakcję z użytkownikiem. Do zalet formatu JSON należy prostota w tworzeniu reprezentacji danych oraz ich interpretacji, zarówno przez ludzi, jak i aplikacje je przetwarzające. Obsługa danych w formacie JSON nie wymaga instalacji dodatkowych rozszerzeń w przypadku wykorzystania tego formatu na potrzeby serwisów internetowych. Dane przesyłane do przeglądarki w postaci kodu zagnieżdżonego w strukturze dokumentu HTML są przetwarzane przez skrypty w języku Java Script, które realizują określone funkcjonalności.

JSON pozwala na definicję typów danych (np. dane liczbowe, ciąg znaków) oraz oferuje własną notację. W założeniu ekspresja danych przyjmuje formę zbioru par atrybut-wartość, która jest zamknięta w nawiasy klamrowe, pełniące również funkcję znaków delimitacyjnych dla obiektów o złożonej strukturze. Przywoływany w tekście przykład danych teleadresowych dla Jana Kowalskiego w uproszczonej postaci formatu JSON przedstawiono poniżej:

```
{
„imię-nazwisko”: „Jan Kowalski”,
„e-mail”: „jan@kowalski.pl”,
„afiliacja”: {
„nazwa”: „Uniwersytet Warszawski”,
„ulica”: „Krakowskie Przedmieście 26/28”,
„miasto”: „Warszawa”,
„państwo”: „Polska”,
„kod-pocztowy”: „00–927”}
}
```

Z punktu widzenia tematu tego artykułu istotny jest kontekst semantyczny odwzorowywany w formatach danych strukturalnych. Chodzi o możliwość formalnej specyfikacji (za pośrednictwem URI przestrzeni nazw) schematów metadanych czy ontologii, wykorzystywanych do ekspresji danych i stosowania semantycznych znaczników. JSON w podstawowej formie nie daje takich możliwości, jak np. RDFa (wykładniki *prefix* oraz *vocab*) czy mikrodane (wykładnik *itemtype*). Problem ten został rozwiązany w postaci rozszerzenia JSON w formie specyfikacji JSON-LD. W przedstawionym przykładzie występują wykładniki elementów metadanych, które wyrażono za pomocą języka naturalnego (np. e-mail, ulica, miasto), co jest zrozumiałe dla człowieka, jednak z punktu widzenia przetwarzania, a szczególnie wymiany danych, powoduje problemy z ich formalną interpretacją.

Prace nad formatem danych JSON-LD doprowadziły do opublikowania w 2012 r. pierwszej wersji roboczej jego specyfikacji. Dwa lata później konsorcjum WWW przyjęło i opublikowało oficjalną rekomendację JSON-LD 1.0 o nazwie A *JSON-based Serialization for Linked Data*.

3.4.1 Poziom semantyczny

Na poziomie koncepcyjnym JSON-LD realizuje model formatu danych JSON, który ma postać grafu skierowanego, którego węzły tworzą elementy danych (np. ciąg znaków, liczba, URI), a krawędzie są wykładnikami atrybutów i relacji (Sporny et al. 2014). Format JSON-LD pełni funkcję serializacji modelu RDF, tak jak to ma miejsce w odniesieniu do RDFa. JSON-LD, tak jak RDFa i mikrodane, pozwala na stosowanie formalnych specyfikacji metadanych i ontologii w celu reprezentacji informacji. JSON-LD pozwala również na jednoczesne stosowanie wielu schematów metadanych w ramach reprezentacji tego samego zasobu informacyjnego. Możliwość odwoływania się do zewnętrznych specyfikacji metadanych została zapewniona poprzez wprowadzenie dodatkowych elementów do składni JSON-LD, co zapewnia zgodność z wytycznymi publikowania danych w modelu Linked Data.

3.4.2 Poziom syntaktyczny

Na poziomie syntaktycznym JSON-LD nie wprowadza żadnych elementów do systemu języka HTML. Zagnieżdżanie danych w formacie JSON czy JSON-LD w dokumencie HTML polega na zastosowaniu skryptu w języku Java Script, który albo pobiera dane z serwera, albo odwołuje się do danych przesłanych wraz ze stroną HTML, które zapisano w sekcji odpowiedzialnej za skrypty (<script>). Jest to więc zgoła inne podejście do zastosowania FDS niż w przypadku mikroformatów, RDFa i mikrodanych, gdzie formaty te są wykorzystywane do zagnieżdżania danych bezpośrednio w treści dokumentu.

JSON-LD jako rozszerzenie JSON wprowadza szereg nowych elementów syntaktycznych, które pozwalają na publikowanie danych w modelu Linked Data wewnątrz struktury dokumentu HTML. Cztery najważniejsze z nich to:

(1) *@context* – element systemu znacznikowego służący do identyfikacji bazowej przestrzeni nazw (URI) schematu metadanych, z którego elementy będą wykorzystywane w opisie;
(2) *@id* – element systemu znacznikowego służący do identyfikacji opisywanego obiektu z wykorzystaniem identyfikatora URI;
(3) *@type* – element systemu znacznikowego służący do kategoryzacji opisywanego obiektu (np. osoba, wydarzenie);
(4) *@vocab* – element systemu znacznikowego służący do deklaracji dodatkowych przestrzeni nazw, które rozszerzają możliwości formalnej ekspresji metadanych.

Poniżej przedstawiono uproszczoną wersję formatu JSON-LD dla przywoływanego wcześniej przykładu danych teleadresowych dla Jana Kowalskiego:

```
{
 «@context»: «http://schema.org/»,
 «@type»: «Person»,
 “name”: “Jan Kowalski”,
 “affiliation”: {
       “@context”: “http://schema.org/”,
       „@type”: „Organization”,
       „name”: „Uniwersytet Warszawski”,
„address”: {
„@type”: „PostalAddress”,
„streetAddress”: „Krakowskie Przedmieście 26/28”,
“addressLocality”: “Warszawa”,
“postalCode”: “00–927”,
“addressCountry”: “Polska”}
}
}
```

Mamy tutaj do czynienia z reprezentacją informacji, której przedmiotem jest Jan Kowalski formalnie skategoryzowany jako instancja klasy człowiek w ontologii schema.org. Tak jak w poprzednich przykładach afiliacja Jana Kowalskiego została odwzorowana za pomocą prostych elementów metadanych (nazwa) oraz obiektu złożonego odwzorowującego dane teleadresowe Uniwersytetu Warszawskiego. Zdefiniowanie kontekstu poprzez użycie deklaracji „@context”: „http://schema.org/” powoduje, że wszystkie zastosowane wykładniki metadanych, chociaż zapisywane w skróconej postaci (np. „name”), są formalnie interpretowane za pośrednictwem URI w zadeklarowanej przestrzeni nazw (np. http://schema.org/name).

Nadanie tożsamości sieciowej osobie Jana Kowalskiego w przytoczonym przykładzie polega na zastosowaniu elementu *@id* i przeformułowaniu deklaracji zawierającej kontekst:

```
{
 „@context”:
{
„@vocab”: „http://schema.org/”
},
 „@id”: „ http://przyklad.pl/JKowalski”,
 „@type”: „Person”,
 „name”: „Jan Kowalski”,

      ...
```

W tym przypadku formalny kontekst semantyczny deklaracji, wyrażonej za pomocą formatu JSON-LD, został wyrażony poprzez zdefiniowanie przedmiotu odniesienia (elementów schematu schema.org) za pośrednictwem wykładnika *„@vocab”*.

Specyfikacja JSON-LD zawiera szczegółową charakterystykę elementów tego formatu danych, które pozwalają na bardzo szczegółową reprezentację danych oraz konstrukcję elementów złożonych. Zasadniczą cechą JSON-LD jest zbieżność koncepcyjna z modelem RDF oraz operowanie URI zarówno dla elementów zastosowanego schematu metadanych, jak i dla obiektów zdefiniowanych na danej stronie internetowej.

3.4.3 Zalety i ograniczenia

Do podstawowych zalet tego formatu danych należy z pewnością zaliczyć jego rozpowszechnienie wśród twórców aplikacji i usług sieciowych oraz prostotę w użyciu. Dane w formacie JSON-LD nie są zapisywane bezpośrednio w treści strony WWW, lecz w osobnej sekcji przeznaczonej dla zagnieżdżania skryptów. W przypadku interpretacji JSON-LD nie jako formatu wymiany danych, lecz jako dodatkowej warstwy semantycznej odwzorowywanej w ramach struktury dokumentu HTML, można to traktować zarówno jako zaletę jak i mankament. W pierwszym ujęciu mamy ewidentne rozdzielenie treści strony WWW od metadanych, w drugim – mamy do czynienia z powielaniem pewnych elementów treści.

3.5. Podsumowanie

Przedstawiona charakterystyka czterech formatów danych strukturalnych daje obraz przede wszystkim sposobu konstrukcji i ekspresji metadanych w ramach systemu języka znacznikowego HTML, kładąc nacisk na rozwiązania syntaktyczne oraz otwartość na implementację istniejących schematów metadanych i ontologii sieciowych, stosowanych w kolekcjach cyfrowych. Mikroformaty były pierwszą próbą wzbogacenia struktury stron WWW o znaczniki semantyczne pozwalające na głębszą strukturyzację zawartości. RDFa jest wynikiem próby przeniesienia koncepcji reprezentacji wiedzy standardu RDF stosowanego w bazach wiedzy na poziom stron WWW, zaś mikrodane wydają się być kompromisem między prostotą mikroformatów, a szczegółową strukturyzacją charakterystyczną dla RDFa. JSON-LD to z kolei „ukłon” konsorcjum WWW w stronę twórców aplikacji sieciowych, dla których format JSON jest naturalnym środowiskiem danych. Tabela 2 zawiera porównanie czterech opisanych formatów danych strukturalnych z uwzględnieniem wybranych cech. Manu Sporny (2015) opracował szczegółowe porównanie mikroformatów, RDFa oraz mikrodanych, zawierające ich pogłębione analizy.

Z punktu widzenia ingerencji w system języka znacznikowego HTML najwięcej nowych elementów syntaktycznych wprowadza RDFa, co niewątpliwie ma wpływ na siłę ekspresji tego FDS. Zarówno mikrodane, RDFa jak i JSON-LD są otwarte na implementację zewnętrznych schematów metadanych i ontologii sieciowych w kontekście doboru narzędzi reprezentacji informacji adekwatnego dla opisywanej kolekcji. W przypadku mikrodanych wiąże się to jednak w praktyce ze stosowaniem specyfikacji schema.org, może nawet nie tyle ze względu na siłę semantyczną tego schematu metadanych, co jego rekomendację przez dostawców wyszukiwarek internetowych i tym samym – dążenie do „lepszej widoczności" tak zaindeksowanych zasobów sieciowych w rezultatach wyszukiwania. Mikroformaty są ściśle powiązane z konkretnymi specyfikacjami metadanych. Zarówno mikrodane, jak i RDFa, pozwalają na odwoływanie się do opisywanych obiektów oraz ich atrybutów za pośrednictwem ich unikalnych URI. Daje to większą możliwość standaryzacji i tym samym gotowość na agregację danych z wielu źródeł.

Tab. 2. Porównanie formatów danych strukturalnych; źródła: Sporny, 2015 oraz badania własne

Cecha	Mikroformaty	Mikrodane	RDFa	JSON-LD
Liczba nowych elementów syntaktycznych HTML	0	5	8	Nie dotyczy
Powiązanie z konkretnym schematem metadanych	Tak	Nie[1]	Nie	Nie
Identyfikacja opisywanego obiektu za pomocą URI	Nie	Tak	Tak	Tak
Formalna kategoryzacja opisywanego zasobu informacyjnego z wykorzystaniem URI	Nie	Tak	Tak	Tak
Możliwość jednoczesnego stosowania wielu schematów metadanych w ramach jednego zestawu deklaracji	Nie	Tak	Tak	Tak
Trudność w implementacji	Niska	Średnia	Wysoka	Niska

[1] Formalnie mikrodane nie są związane z konkretnym schematem metadanych, ale w praktyce ich stosowanie wiąże się z użyciem schema.org

Trudność w implementacji odnosi się nie tyle do aspektu technicznego, co do pragmatyki, czyli zastosowania w procesie adnotowania treści stron internetowych przez ich twórców. W takim ujęciu RDFa jest wskazywane jako stosunkowo trudne we wdrożeniu, przy średniej ocenie mikrodanych i prostocie mikroformatów oraz JSON-LD.

Przedstawione FDS można również analizować z punktu widzenia ich genezy. Mikroformaty to bezsprzecznie inicjatywa twórców serwisów internetowych oraz rozwijającej się ówcześnie blogosfery. RDFa to pomysł na uproszczenie tworzenia baz wiedzy z wykorzystaniem RDF na rzecz strukturyzacji bezpośrednio treści stron internetowych i próba implementacji technologii semantycznych do projektowania serwisów WWW. JSON-LD to z kolei odpowiedź konsorcjum WWW na potrzeby twórców aplikacji sieciowych. W końcu mikrodane to głos twórców języka HTML w dyskusji nad adekwatnym formatem danych strukturalnych.

4. Zakres wykorzystania formatów danych strukturalnych w World Wide Web

Z dotychczas przedstawionych rozważań wyłania się obraz czterech formatów danych przeznaczonych do formalnej strukturyzacji treści stron WWW, z których mikrodane, RDFa i JSON-LD operują zbliżonymi środkami ekspresji. Obecność wielu standardów jest weryfikowana w praktyce przez ich realne wykorzystanie przez społeczność, do której są skierowane. Na wybór danego rozwiązania oprócz jego efektywności ma wpływ wiele czynników, do których na pewno można zaliczyć poziom ich trudności oraz stosunek poniesionych nakładów do realnych korzyści wynikających z ich zastosowania. Richard K. Bergman (2011) stwierdził, że twórcy serwisów WWW będą wybierali formaty, które są łatwe do zrozumienia, posiadają kompletną dokumentację i są tym samym łatwe do implementacji. Zdaniem Bergmana takim wyborem są mikrodane. Chociaż przywoływany pogląd pochodzi z 2011 r., to zaprezentowane poniżej wyniki badań dotyczących zakresu wykorzystania FDS wskazują na jego aktualność.

Celem tej części artykułu jest określenie zakresu wykorzystania mikroformatów, RDFa, mikrodanych oraz JSON-LD w zasobach WWW w celu strukturyzacji treści stron internetowych. Przedmiotem przeprowadzonych badań była dokumentacja zbioru danych badawczych, która powstała w ramach projektu Web Data Commons[12]. Projekt ten został zainicjowany w 2012 r. w ramach współpracy między niemieckimi instytutami badawczymi – Wolnym Uniwersytetem Berlina (Freie Universität Berlin) oraz politechniką Karlsruher Institut für Technologie. Celem projektu było cykliczne ekstrahowanie ze zbioru Common Crawl[13] (największego i publicznie dostępnego korpusu zasobów sieciowych) informacji o stronach internetowych, w których ich twórcy wykorzystali FDS jako narzędzia formalnej strukturyzacji informacji. Do projektu włączono dodatkowo graf hiperłączy oraz kolekcję tabel, które derywowano z dokumentów HTML. Analizy zakresu wykorzystania wspomnianych formatów danych są prowadzone co roku począwszy od 2009 r. (z wyłączeniem 2011 r.). Od 2015 r. badania obejmują również format JSON-LD. Ostatnie badanie przeprowadzone w ramach projektu Web Data Commons zawiera dane reprezentujące stan na październik 2016 r.

Tab. 3. Ogólne statystyki badania Web Data Commons z 2016 r.

Zmienna	Wartość
Liczba przetworzonych URL	3 181 199 447
Liczba URL zawierających deklaracje	1 242 727 852
Liczba przetworzonych adresów domenowych (PLD)	34 076 469
Liczba domen (PLD), których zasoby zawierają dane strukturalne	5 638 796
Liczba zdefiniowanych obiektów	9 590 731 005
Liczba deklaracji	44 242 655 138

Metodologia projektu Web Data Commons zakłada automatyczną rejestrację informacji nie tylko o zastosowanym formacie danych strukturalnych na danej stronie WWW, ale

[12] http://webdatacommons.org/
[13] http://commoncrawl.org/

również ekstrahowanie całego zbioru danych ustrukturyzowanych z określonego źródła sieciowego[14]. Według danych z października 2016 r. przeanalizowano ponad 3.2 mld adresów URL z czego ponad 1.2 mld stron internetowych zawierały formalnie specyfikowane deklaracje świadczące o obecności danych ustrukturyzowanych (Tab. 3). Łącznie przeanalizowano ponad 54 terabajty danych.

Przez deklarację rozumie się tutaj elementarną jednostkę wypowiedzi z wykorzystaniem metadanych w układzie <obiekt>-<atrybut>-<wartość>, którą w badaniu określa się mianem trójki (ang. *triple*), co nawiązuje do modelu danych RDF. Z uogólnionych danych przedstawionych w tabeli 3 wynika, że dane ustrukturyzowane były obecne prawie w 40% stron internetowych poddanych badaniu. Na potrzeby metodologii badań wprowadzono pojęcie domeny poziomu płatnego (ang. *Pay-Level Domain*; PLD), przez które rozumie się subdomenę wyodrębnioną w ramach domeny najwyższego poziomu (np. .pl, .gov, .edu)[15]. Jest to więc indywidualna nazwa domenowa, do której używania należy nabyć prawa. PLD wykorzystano jako zmienną do analiz przede wszystkim ilościowych, tzn. do określenia zakresu wykorzystania danych ustrukturyzowanych w ramach określonego serwisu internetowego. Z danych przedstawionych w tabeli 3 wynika, że dane ustrukturyzowane były obecne w 16% źródeł internetowych. Przez obiekt zdefiniowany (ang. *typed entity*) rozumie się obiekt, który został formalnie skategoryzowany, tzn. istnieje deklaracja (trójka), która zawiera informacje o typie obiektu, który on reprezentuje (np. wpis na blogu, przepis kulinarny, osoba, książka, itp.)[16]. Z udostępnionych danych wynika więc, że na jednej stronie internetowej, zawierającej dane ustrukturyzowane obecne było średnio 35 deklaracji wskazujących na własności średnio osiem zdefiniowanych obiektów.

Tab. 4. Porównanie zmian w wartościach zmiennych analizowanych w badaniach z lat 2012–2015

	2013	**2014**	**2015**	**2016**
Odsetek przetworzonych URL	-25.98	-9.47	-12.10	+44.34
Odsetek URL zawierających deklaracje	+58.64	+5.87	-12.68	+56.43
Odsetek przetworzonych domen	-68.40	+22.11	-8.04	+57.71
Odsetek domen zawierających deklaracje	-22.15	+52.95	+0.08	+51.68
Odsetek zdefiniowanych obiektów	+135.42	+29.35	+10.72	+36.32
Odsetek deklaracji	+134.55	+18.81	+19.00	+44.90

W tabeli 4 zestawiono aktualne dane dotyczące próby badawczej i zakresu wykorzystania omawianych formatów danych ze stanem z lat ubiegłych. Tabela prezentuje zmiany w wartościach omawianych wcześniej zmiennych, które zaszły w stosunku do roku poprzedniego.

[14] Dane badawcze z listopada 2016 r. to zbiór ponad 9000 plików o łącznej wielkości 967 gigabajtów. Tym samym istnieje możliwość pobrania całego zbioru danych i przeprowadzania na nim dalszych analiz. Dane badawcze z października 2016 r. są dostępne pod adresem http://webdatacommons.org/structureddata/2016–10/stats/how_to_get_the_data.html

[15] Np. stronę internetową http://www.przyklad.com.pl/zasoby/strona1.html oraz http://www.przyklad.com.pl/zasoby/strona5.html kwalifikowano do jednej PLD – http://przyklad.com.pl

[16] W przypadku mikroformatów jest to wystąpienie atrybutu class=typ, RDFa – typeof=typ, mikrodane: itemtype=typ, JSON-LD – @type=typ.

W ciągu ostatnich lat analizowane były zbiory o różnej wielkości; można zauważyć ciągły wzrost liczby zarówno deklaracji, jak i formalnie zdefiniowanych obiektów, o których informacje są zagnieżdżane na stronach WWW. Jest to szczególnie widoczne dla 2016 r.

Największy wzrost odsetka stron WWW zawierających formalnie specyfikowane deklaracje na temat ich treści w stosunku do roku wcześniejszego zaobserwowano w 2013 r. (ponad 58%), przy jednocześnie mniejszej próbie badawczej (spadek o ponad 25%). W tym roku odnotowano również wzrost o ponad 130% zarówno odsetka deklaracji, jak i formalnie zdefiniowanych elementów w treści stron WWW. W latach 2014–2015 widoczny jest nieznaczny wzrost zainteresowania publikowaniem danych ustrukturyzowanych w treści stron WWW, natomiast dane za 2016 r. pokazują duży skok w odniesieniu do wszystkich zmiennych.

Tab. 5. Średnia liczba zdefiniowanych obiektów i deklaracji w treści stron internetowych wykorzystujących formaty danych strukturalnych z lat 2012–2016

	2012	**2013**	**2014**	**2015**	**2016**
Średnia liczba obiektów na stronie	5	7	9	11	8
Średnia liczba deklaracji na stronie	20	29	33	45	36

Z punktu widzenia ilości informacji specyfikowanych za pomocą omawianych formatów danych (Tab. 5) od 2012 r. utrzymuje się tendencja wzrostowa. Dotyczy to zarówno średniej liczby zdefiniowanych obiektów, jak i liczby deklaracji. Spadek w 2016 r. nie oddaje tego stanu, ponieważ zbiór badanych stron WWW zwiększył się prawie o połowę.

Tab. 6. Wyniki badania Web Data Commons z października 2016[17]

Format	**PLD**	**URL**	**Obiekty zdefiniowane**	**Deklaracje**
html-microdata	2 537 539	901 118 191	6 872 341 887	34 637 805 559
html-embedded-jsonld	2 116 755	111 411 049	385 731 201	1 880 721 886
html-mf-hcard[1]	1 668 039	159 748 255	1 614 688 960	4 600 477 456
html-rdfa	93 883	311 533 110	511 555 208	2 216 933 416
html-mf-xfn	195 595	24 242 546	48 011 285	300 764 344
html-mf-adr	188 755	27 697 569	80 039 476	259 718 235
html-mf-geo	238	6 151 013	14 644 289	25 733 274
html-mf-hcalendar	22 313	3 450 075	33 962 568	177 931 362
html-mf-hreview	16 984	4 551 011	13 680 480	79 631 745
html-mf-hlisting	471	37 418	9 578 853	36 521 676
html-mf-hrecipe	2 923	755 544	5 695 917	24 347 685
html-mf2-h-adr	1 415	1 362	26 659	726 401
html-mf-hresume	168	2 961	7 106	22 262
html-mf-species	95	170 516	527 185	1 319 837

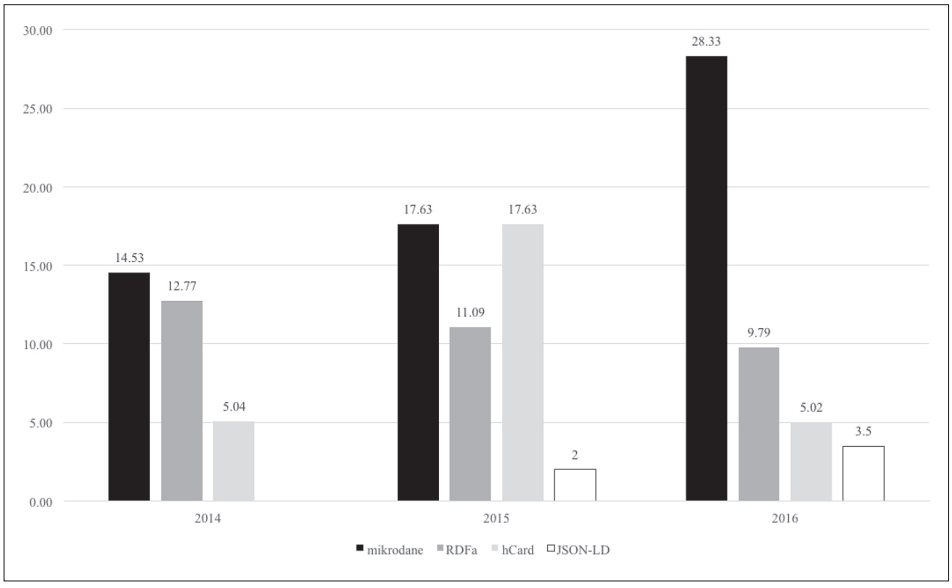
[1] html-mf-hcard: format dla danych kontaktowych; html-mf-xfn: format danych do specyfikacji relacji interpersonalnych; html-mf-adr oraz html-mf2-h-adr: formaty danych teleadresowych; html-mf-geo: format danych dla lokalizacji geograficznej; html-mf-hcalendar: format danych dla wydarzeń; html-mf--hreview: format danych dla recenzji; html-mf-hlisting: format danych dla ofert handlowych; html-mf--hrecipe: format danych dla przepisów kulinarnych; html-mf-hresume: format danych dla Curriculum Vitae; html-mf-species: format danych dla reprezentacji informacji z obszaru taksonomii biologicznych.

[17] http://www.webdatacommons.org/structureddata/2016–10/stats/stats.html

Aktualny obraz zakresu strukturyzacji informacji wewnątrz dokumentów HTML zawierają dane z października 2016 r. W tabeli 6 przedstawiono rozkład wartości omawianych wcześniej zmiennych w odniesieniu do mikroformatów (html-mf-xxx), mikrodanych (html--microdata), RDFa (html-rdfa) oraz formatu JSON-LD (html-embedded-jsonld). Z punktu widzenia liczby stron internetowych (URL), w których treści obecne są te formaty największy odsetek wskazań odnosi się do zastosowania specyfikacji mikrodanych (58%) i RDFa (20%). Pozostałe ok. 15% wskazań dotyczy poszczególnych specyfikacji mikroformatów, a ok. 7% wskazań – stron WWW zawierających dane strukturalne w formacie JSON-LD.

Z punktu widzenia PLD największy odsetek odnotowano jednocześnie dla mikrodanych (31.3%) oraz mikroformatów (31.2%). Standard RDFa jest obecny tylko w ok. 15% serwisów internetowych. Z punktu widzenia liczby deklaracji, a więc ilości informacji formalnie specyfikowanych w strukturze stron internetowych, w ponad połowie przypadków (54.2%) są to deklaracje z wykorzystaniem standardu mikrodane. W ok. 37 % przypadków metadane wyrażone zostały za pomocą mikroformatów, a w ok. 6% – w standardzie RDFa.

Rysunek 2 zawiera zestawienie danych wskazujących na realne wykorzystanie FDS w latach 2014–2016. Do analiz wybrano mikrodane, RDFa, JSON-LD oraz najczęściej stosowaną specyfikację mikroformatów – hCard (zob. Tab. 6). Przez realne wykorzystanie FDS rozumie się tutaj stosunek liczby adresów URL, w których zidentyfikowano wystąpienie danego FDS do całego zbioru adresów URL, który został poddany badaniu w danym roku.



Rys. 2. Wykorzystanie FDS w latach 2014–2016

Z danych przedstawionych na rysunku 2 wynika, że w zbiorze danych badawczych można zaobserwować wzrost liczby implementacji mikrodanych w ostatnich trzech latach. Prawie co trzecia strona WWW w zbiorze danych badawczych za 2016 r. zawierała formatowanie z wykorzystaniem mikrodanych. Format RDFa w kolejnych latach był stosowany na podobnym poziomie ok. 10%, natomiast implementacja mikroformatu hCard

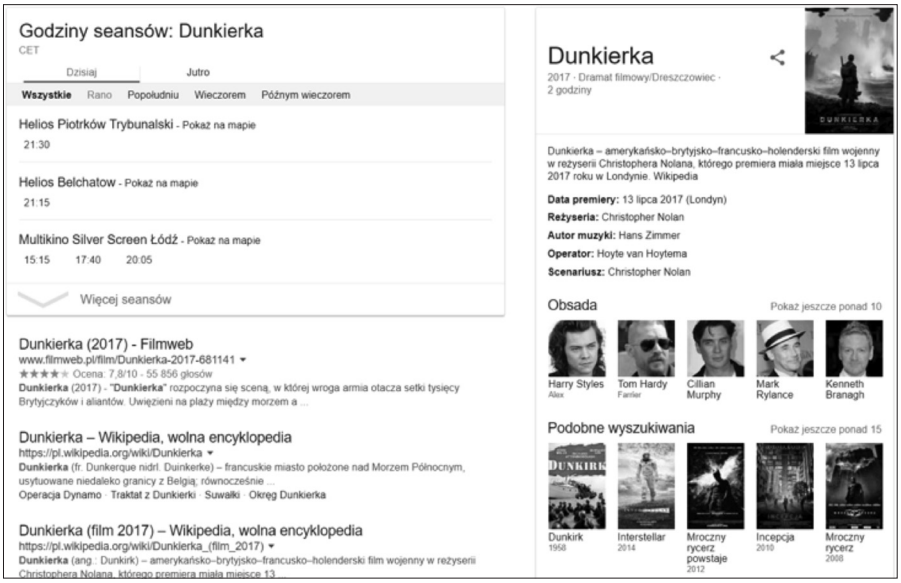
wykazuje wahania i wzrost w 2015, a następnie 50% spadek w kolejnym roku. Można zatem postawić więc pytanie, czy spadek ten miał wpływ na wzrost zastosowania mikrodanych i przyjęcie przez twórców serwisów internetowych tego formatu za podstawowy sposób strukturyzacji treści stron internetowych. W zaprezentowanych analizach widoczne jest również niewielkie zainteresowanie implementacją formatu JSON-LD, które odnotowano w 2015 r., ale wartość tej zmiennej jest stosunkowo mała.

5. Zakończenie

Formaty danych strukturalnych przeznaczone do reprezentacji treści stron internetowych zapewniają nowe możliwości strukturyzacji treści w ramach systemu znaczników języka HTML i tym samym rozszerzają zakres możliwych funkcjonalności mechanizmów wyszukiwawczych, zarówno implementowanych wewnątrz serwisów WWW, jak i wyszukiwarek internetowych o zasięgu globalnym. Taki sposób reprezentacji informacji w porównaniu z metodami reprezentacji wiedzy w Sieci Semantycznej może być określany mianem „płytkiej semantyki" (zob. Hitzler et al., 2012) i wpisuje się w koncepcję wyszukiwania semantycznego (ang. *semantic search; entity search*). „Płytki" charakter semantyki odwzorowywanej za pośrednictwem FDS polega tutaj na zapewnieniu modelu danych i środków syntaktycznych do tworzenia stosunkowo prostych deklaracji na temat faktów zapisanych w treści stron internetowych, z wykorzystaniem formalnie specyfikowanych schematów metadanych oraz elementów ontologii sieciowych. Oznacza to przyjęcie określonego konsensusu w procesie reprezentacji informacji, który jest zdeterminowany pragmatycznie. Znajdująca się na drugim biegunie „głęboka semantyka" jest metaforą sieciowych baz wiedzy operujących ontologiami, gdzie zastosowane formaty danych (np. RDF, OWL) pozwalają zarówno na szczegółową reprezentację informacji, jak i operowanie językiem logiki w celu wnioskowania nowych faktów.

Można więc przyjąć, że implementacja FDS ma na celu zwiększenie efektywności wyszukiwania zasobów sieciowych przez zastosowanie prostych technologii semantycznych, których celem jest formalna identyfikacja i kategoryzacja elementów treści stron WWW oraz odwzorowanie podstawowych faktów na ich temat z wykorzystaniem środków syntaktycznych danego FDS. Taki sposób reprezentacji informacji w zasobach WWW można również interpretować w kontekście semantycznych adnotacji (ang. *semantic annotation*), przez które rozumie się zarówno proces dodawania automatycznie przetwarzalnych metadanych w treści dokumentu, jak i jego efekt, czyli zbiór formalnie specyfikowanych faktów ekstrahowanych z jego treści (Oren et al. 2006). Na wybór określonego formatu, oprócz przytoczonych wcześniej kryteriów, powinna mieć wpływ również odpowiedź na pytanie, czy jest on wspierany przez globalne wyszukiwarki internetowe. Obecność FDS w zasobach WWW można zaobserwować, np. w postaci formy prezentacji rezultatów wyszukiwania m.in. w wyszukiwarkach Bing oraz Google. W przypadku Google mamy do czynienia z generowaniem listy adresów zasobów sieciowych spełniających kryteria wyszukiwania oraz często prezentacją zagregowanych faktów na temat przedmiotu wyszukiwania, które prezentowane są bezpośrednio pod nazwą zasobu sieciowego w rezultatach wyszukiwania oraz w panelu bocznym w interfejsie graficznym przeglądarki internetowej. Wykorzystanie globalnego identyfikatora sieciowego pozwala bowiem na agregację danych na temat

danego obiektu w rozproszonym środowisku sieciowym World Wide Web. Taką praktykę stosują dostawcy wspomnianych wyszukiwarek internetowych, co w przypadku Google przejawia się generowaniem wyników rozszerzonych, czy to w postaci dodatkowych informacji o wyszukanych zasobach sieciowych prezentowanych bezpośrednio w rezultatach wyszukiwania (tzw. *rich snippets*), czy też w postaci ustrukturyzowanych metadanych na temat obiektu zidentyfikowanego przez algorytm Google jako przedmiot wyszukiwania, które prezentowane są m.in. w panelu bocznym (Rys. 3).



Rys. 3. Przykład rezultatów wyszukiwania w wyszukiwarce Google.pl dla zapytania „Dunkierka"

Na rysunku 3 przedstawiono fragment prezentacji rezultatów wyszukiwania w wyszukiwarce Google.pl dla zapytania „Dunkierka". Wykaz wyników oraz elementów ustrukturyzowanych prezentowanych użytkownikowi pozwala na postawienie wniosku, że mechanizm tej wyszukiwarki na podstawie zastosowanych algorytmów przetwarzania własnej bazy, zidentyfikował przedmiot wyszukiwania jako film pt. „Dunkierka". W panelu bocznym zaprezentowano wybrane fakty na temat tego filmu oraz dodatkowo na podstawie geolokalizacji komputera, z którego wysłano zapytanie, wskazano na kina w pobliżu, w których wyświetlany jest ten film. Zaprezentowane informacje o charakterze faktograficznym są wynikiem agregacji danych, które przeprowadził mechanizm indeksujący i z dużą pewnością można założyć, że informacje te są zapisane z wykorzystaniem formatów danych strukturalnych[18].

Publikowanie informacji w WWW z wykorzystaniem formatów danych strukturalnych z pewnością stanowi również szansę dla bibliotek i repozytoriów cyfrowych z jednej strony,

[18] Obecność danych strukturalnych można zweryfikować za pomocą wielu aplikacji dostępnych online. Jedną z nich jest „Narzędzie do testowania uporządkowanych danych" oferowane przez Google. Zob. np. test zawartości rekordu dla filmu pt. „Dunkierka" w serwisie Fimweb.pl – https://search.google.com/structured-data/testing-tool/u/0/#url=http%3A%2F%2Fwww.filmweb.pl%2Ffilm%2FDunkierka-2017–681141

na lepszą „widoczność" tych kolekcji bezpośrednio w rezultatach wyszukiwania w wyszukiwarkach globalnych, a z drugiej, na włączenie ich do globalnej sieci danych (ang. *Web of Data*) poprzez wykorzystanie stosunkowo prostych technologii semantycznych. Wybór konkretnego formatu danych strukturalnych ma tutaj oczywiście znaczenie, ale wydaje się że mając na uwadze zachowania i kompetencje informacyjne użytkowników, za ważniejszą uznać trzeba świadomość konieczności implementacji takich technologii.

Bibliografia

Bergman, M. (2011). Structured Web Gets Massive Boost [online]. AI3[20.08. 2017], http://www.mkbergman.com/962/structured-web-gets-massive-boost/

Berners-Lee, T. (1994). The World-Wide Web. *Communications of the ACM1*, 37(8), 792–799.

Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web [online]. *Scientific American* (May 17), [20.08.2017], http://www.scientificamerican.com/article.cfm?id=the-semantic-web

Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems* [online], 5(3), [20.08.2017], http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

Bizer, C., Mendes, P. N., Jentzsch, A. (2012). Topology of the Web of Data. In: R. De Virgilio, F. Guerra & Y. Velegrakis (eds.), *Semantic Search over the Web* (3–29). Berlin, Heidelberg: Springer , http://doi.org/10.1007/978–3-642–25008

Guha, R. V., Brickley, D., Macbeth, S. (2015). Schema.org: Evolution of Structured Data on the Web. *ACMQUEUE* {online], 9(13). [20.08.2017], http://queue.acm.org/detail.cfm?id=2857276

Herman, I., Adida, B., Sporny, M. (2015). *RDFa 1.1 Primer – Third Edition. Rich Structured Data Markup for Web Documents* [online]. W3C [20.08.2017], https://www.w3.org/TR/rdfa-primer/

Hitzler, P., Janowicz, K., Berg-Cross, G., Sheth, A., Finin, T., Cru, I. (2012). *Semantic Aspects of EarthCube* [online]. EarthCube [20.08.2017], https://www.earthcube.org/document/2012/semantic--aspects-earthcube

McCathie Nevile, C., Brickley, D. (2017). *HTML Microdata* [online]. *W3C Working Draft 26 June 2017* , [20.08.2017] https://www.w3.org/TR/microdata/

Oren, E., Möller, K. H., Scerri, S., Handschuh, S., Sintek, M. (2006). *What Are Semantic Annotations?* [online]. Prof. Siegfried Handschuh [20.08.2017] http://www.siegfried-handschuh.net/pub/2006/whatissemannot2006.pdf

Ronallo, J. (2012). HTML5 Microdata and Schema.org. *The Code4Lib Journal* [online], (16). [20.08.2017], http://journal.code4lib.org/articles/6400

Sikos, L. F. (2015). *Mastering Structured Data on the Semantic Web: From HTML5 Microdata to Linked Open Data*. Berkeley, CA: Apress.

Sporny, M. (2015). An Uber-Comparison of RDFa, Microdata and Microformats [online]. Internet Archive Wayback Machine [20.08.2017], https://web.archive.org/web/20160329022408/http://manu.sporny.org/2011/uber-comparison-rdfa-md-uf/

Sporny, M., Kellogg, G., Lanthaler, M. (2014). JSON-LD 1.0. *A JSON-based Serialization for Linked Data* [online]. W3C [20.08.2017], https://www.w3.org/TR/json-ld/

Tomberg, V., Laanpere, M. (2009). RDFa versus Microformats: Exploring the Potential for Semantic Interoperability of Mash-up Personal Learning Environments [online]. In: F. Wild, M. Kalz, M. Palmer & D. Muller (eds.), *Mash-Up Personal Learning Environments. Proc. of the 2nd Workshop MUPPLE'09, Nice, France, September 29, 2009, CEUR* (102–109). CEUR. [20.08.2017], http://ceur-ws.org/Vol-506

Wetherill, B. (2014). RDFa and Microdata. *Library Philosophy and Practice (E-Journal)* [online], 1151, 19. [20.08.2017], http://digitalcommons.unl.edu/libphilprac/1151/

Structured Data Formats for World Wide Web

Abstract

Purpose/Thesis: The aim of this paper is the analysis and comparison of data formats for the content representation of Web pages embedded in HTML structure. The subjects of investigation are four structured data formats: microformats, RDFa, microdata and JSON-LD and their implementation on the Web.
Approach/Methods: The starting points for the investigation are two statements. The first one is that the separation between content and presentation layer is one of important features of the World Wide Web and the second refers to the fact that the structure level of Web content is the determining factor for the types of functionality that search engines can provide. These two approaches offer the background for the concept of structured data formats aimed at the formal representation of Web page content using HTML language system. The subjects were selected based on the scope of their implementation on the Web. The analysis was based on the assumption that structured data formats may be investigated from the metadata perspective with the premise that the annotation act is not made on the document level but is related to the facts that constitute the content. The study on structured data formats is based on semantic and syntactic analysis of their features.
Results and conclusions: Structured data formats for the content representation of Web pages provide new methods for knowledge representation by means of HTML language and thus extend the functionalities of both locally implemented and global search mechanisms. The results of the survey conducted in the years 2012–2016 indicate the growth of the interest in the semantic representation of Web pages. RDFa represents a high level of specificity but microdata seem to be the consensus between the desired expressiveness and the ease of implementation, confirmed withthe results of Web Data Commons project.
Originality/Value: The comparison of selected features of four structured data formats offers a clear picture of their capability for deep content annotations with metadata schemes and ontologies. The results from Web Data Commons project for the period 2014–2016 indicate that microdata and schema.org will play an important role in the domain of applying semantic technologies to Web page content representation.

Keywords

JSON-LD. Linked Data. Metadata. Microdata. Microformats. RDFa. Schema.org. Semantic Web.

Dr MARCIN ROSZKOWSKI jest adiunktem w Katedrze Informatologii na Wydziale Dziennikarstwa, Informacji i Bibliologii Uniwersytetu Warszawskiego. Jest członkiem International Society for Knowledge Organization oraz Komitetu ds. ontologii w projekcie DBpedia. Jego zainteresowania naukowe obejmują problematykę organizacji wiedzy i reprezentacji informacji w środowisku sieciowym, ze szczególnym uwzględnieniem modelowania konceptualnego systemów informacyjnych oraz metadanych i ontologii sieciowych. Najważniejsze publikacje: B. Sosińska-Kalata, M. Roszkowski (2016). Organizacja informacji i wiedzy. W: W. Babik (red.), Nauka o informacji (305–358). Warszawa: Wydaw. SBP; M. Roszkowski, W. Mustafa El Hadi (2016). The Role of Digital Libraries as Virtual Research Environments for the Digital Humanities. In: J. A. C. Guimarães, S. Oliveira Milani, & V. Dodebei (eds.), Advances in Knowledge Organization (Vol. 15). Ergon Verlag, 392–402; M. Roszkowski (2016). Kartoteka haseł wzorcowych jako usługa sieciowa – automatyczna identyfikacja nazw osobowych z wykorzystaniem kartoteki VIAF. W: J. Woźniak-Kasperek & J. Franke (red.), Bibliografia – teoria, praktyka, dydaktyka (203–222). Warszawa: Wydaw. SBP.

Kontakt z autorem
m.roszkowski@uw.edu.pl
Katedra Informatologii
Wydział Dziennikarstwa, Informacji i Bibliologii
Uniwersytet Warszawski
ul. Nowy Świat 69
00–046 Warszawa