Schema.org

Dalibor Bačovský

28. listopadu 2018

1 Úvod

Schema.org je komunita s cílem vytvořit, spravovat a propagovat schémata pro strukturovaná data na Internetu, webových stránkách, emailových zprávách a více. Schema.org slovník popisuje jak zapisovat strukturované data na webu tak, aby byla jednoduše strojově zpracovatelná. Je propagováno několika velkými vyhledávači na internetu, které informace popsané pomocí Schema.org slovníku používají ve svých výsledcích.

2 Strukturovaná data na webu

Web je rozsáhlým zdrojem strukturovaných informací, avšak ty jsou zapsané v HTML kódu a v minulosti nesdílely žádný společný design, na rozdíl od strukturovaných databází, na jejímž základu jsou stránky často generovány. Tyto databáze však nebyly často dostupné jinak, než přes HTML webové rozhraní. Služby a aplikace, které chtěly strukturovaná data z webu použít, musely tvořit nástroje na extrakci dat (tzv. scrapery) přímo z HTML kódu, které se mohly při jakékoliv změně rozbít, nebo musely být dostatečně komplexní a počítat s nejistým vzhledem informací. [1, 6]

Obrázek 1: Nahoře výsledek v Google, dole HTML kód, který obsahuje strukturované informace zapsané pomocí Schema.org.

3 Popis

Slovník Schema.org slouží k zápisu strukturovaných informací na webu, v emailových zprávách a podobně. Tento slovník popisuje typy (třídy) a jejich vlastnosti. Typy jsou organizovány hierarchicky, a každý typ může mít více rodičů, avšak většina má pouze jednoho. Třídy, ty jsou zároveň i typem, umožňují popis mnoha entit jako jsou lidé, místa, události, produkty, nabídky a další. Dále definuje základní datové typy jako jsou čísla, řetězce, boolean hodnoty, data apod. Každý typ má několik vlastností a každá vlastnost je zároveň i typem, což jim dává stejnou hierarchii. [9, 6]

Vlastnosti mohou mít jeden nebo více typů ve své doméně nebo rozsahu hodnot, tím se vyhne nutnosti definovat zbytečné nadtypy, jako je tomu například v RDF Schema. Více typů v doméně umožňuje sdílet vlastnosti mezi vzájemně nesouvisejícími typy, např. třídy *Brand*, *Place*, *Service* obsahují vlastnost *logo*. Více typů v rozsahu hodnot dovoluje vyhnout se nutnosti definovat jejich společného nadbytečného nadtypu, např. vlastnost *licence* dovoluje hodnoty typu *CreativeWork* a *URL*. [6, 4]

```
<script type="application/ld+json">
{
    "@context": "http://schema.org",
    "@type": "Book",
    "name": "Split Second",
    "numberOfPages": 364,
    "isbn": "978-1517153151",
    "author": {
        "@type": "Person",
        "name": "Douglas E. Richards",
        "birthDate": "1962-5-7"
    },
    "aggregateRating": {
        "@type": "AggregateRating",
        "ratingValue": "4.2",
        "ratingCount": 6047
    },
    "genre": "Science Fiction",
    "datePublished": "2015-8-31"
}
</script>
```

Zdrojový kód 1: Použití Scheme.org pro popis knižky. Data json z [10]

.

4 Historie

Schema.org bylo vytvořeno v roce 2011 ve spolupráce společností Bing, Google a Yahoo, k těm se později toho roku připojil Yandex. Cílem tohoto projektu bylo sjednotit zápis strukturovaných informací na webu, za účelem rozšíření informací, která nabízejí vyhledávače ve svých výsledcích. Před vznikem Schema.org implementovaly vyhledávače různé slovníky umožňující tvůrcům webových stránek přidat strukturované informace na svoje weby. Slovník, který bych dosáhl jisté popularity před Schema.org, byl FOAF[1] (Friend of a Friend), uvedený v roce 2000. [6, 2]

5 Formáty

Formáty propagované Schema.org komunitou jsou RDFa[2], Microdata[3], JSON-LD[4], z nichž je RDFa nejpoužívanější[5] pro záznam strukturovaných informací na webu. Protože Schema.org je pouze slovník, je ho možné použít s jakýmkoliv formátem či syntaxí pro strukturovaná data. Zde budou dále popsané tři propagované formáty. [5]

5.1 RDFa

Je to standard W3C[6] pro označování a zápis strukturovaných dat do (X)HTML kódu. Původně definováno pouze pro XHTML, ve své novější verzi podporuje jakýkoliv jazyk založený na XML, jako je například SVG. V současnosti existují dvě verze: RDFa Core a její zjednodušená verze RDFa Lite. Lite verze se skládá pouze ze základních 5 atributů: *vocab*, *typeof*, *property*, *resource* a *prefix*. Definice použitého slovníku je provedena atributem *vocab*, typ (třída) informací je zaznačena v atributu *typeof* a vlastnosti uvádí atribut *property*. Plná Core verze podporuje mnoho více možností, jednou z nich je atribut *content*, který obsahuje strojově čitelné informace, zatímco obsah elementu může obsahovat informace lidsky čitelné. [7, 11]

```
<div vocab="https://schema.org/" typeof="Event">
    <span property="name">Example Event</span>
    <span property="location">Brno</span>
    <span property="startDate" content="2018-11-28">
        Begins on 28th November 2018.
    </span>
</div>
```

[1] http://www.foaf-project.org
[2] https://rdfa.info
[3] https://w3.org/TR/microdata
[4] https://json-ld.org
[5] https://w3techs.com/technologies/overview/structured_data/all
[6] konsorcium utvářející standardy pro World Wide Web

Zdrojový kód 2: Příklad data zapsaných pomocí RDFa.

5.2 Microdata

Standard WHATWG[7], podobně jako RDFa, umožňuje vkládání a označování strojové čitelných dat přímo do (X)HTML kódu. Každý typ (třída) informací musí být uveden atributem *itemscope* a *itemtype*, který obsahuje URL na definici typu. Vlastnosti třídy jsou dále označovány atributem *itemprop*. Na rozdíl od RDFa neumožňuje přidávat strojově zpracovatelná data přímo k lidsky čitelným, musí se proto využít tagu *meta*, který nese strojově čitelnou informaci. [3]

```
<div itemscope itemtype="https://schema.org/Event">
    <span itemprop="name">Example Event</span>
    <span itemprop="location">Brno</span>
    <meta itemprop="startDate" content="2018-11-28">
    <span>Begins on 28th November 2018.</span>
</div>
```

Zdrojový kód 3: Příklad data zapsaných pomocí Microdata.

5.3 JSON-LD

Formát založený na JSON (JavaScript Object Notation) pro propojená data (Linked Data). Syntax vychází z JSON pro snadnější integraci se systémy, které již JSON podporují. [8]

```
<script type="application/ld+json">
{
    "@context": "http://schema.org",
    "@type": "Event",
    "name": "Example Event",
    "location": "Brno",
    "startDate": "2018-11-28"
}
</script>
```

Zdrojový kód 4: Příklad data zapsaných JSON-LD.

6 Závěr

Schema.org je užitečný pro definování společného zápisu strukturovaných dat napříč weby, která mohou sloužit pro mnohé účely a aplikace. Díky tomu je

[7] skupina složená hlavně ze společností produkující webové prohlížeče, které nebyly spokojeny s přístupem W3C

možné strojově číst data napříč weby, bez nutnosti specifického nástroje pro pouze hrstku vybraných zdrojů.

Reference

[1] Michael J. Cafarella, Alon Halevy a Jayant Madhavan. „Structured Data on the Web“. In: *Commun. ACM* 54.2 (2011-02), s. 72–79. ISSN: 0001-0782. DOI: 10.1145/1897816.1897839. URL: http://doi.acm.org/10.1145/1897816.1897839.

[2] *Data Model - schema.org*. 2011-11-04. URL: http://blog.schema.org/2011/11/yandex-now-supports-schemaorg-markup.html.

[3] *Data Model - schema.org*. 2011-11-04. URL: http://blog.schema.org/2011/11/yandex-now-supports-schemaorg-markup.html.

[4] *Data Model - schema.org*. URL: https://schema.org/docs/datamodel.html.

[5] *Getting Started - schema.org*. URL: https://schema.org/docs/gs.html.

[6] R. V. Guha, Dan Brickley a Steve MacBeth. „Schema.Org: Evolution of Structured Data on the Web“. In: *Queue* 13.9 (2015-11), 10:10–10:37. ISSN: 1542-7730. DOI: 10.1145/2857274.2857276. URL: http://doi.acm.org/10.1145/2857274.2857276.

[7] Ivan Herman, Mark Birbeck, Ben Adida a Manu Sporny. *RDFa 1.1 Primer - Third Edition*. W3C Note. http://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/. W3C, 2015-03.

[8] Gregg Kellogg. *JSON-LD 1.1*. W3C Working Draft. https://www.w3.org/TR/2018/WD-json-ld11-20181011/. W3C, 2018-10.

[9] *Schema.org*. URL: https://schema.org/.

[10] *Split Second: Douglas E. Richards: 9781517153151: Amazon.com: Books*. URL: https://www.amazon.com/Split-Second-Douglas-Richards/dp/1517153158/.

[11] Manu Sporny. *RDFa Lite 1.1 - Second Edition*. W3C Recommendation. http://www.w3.org/TR/2015/REC-rdfa-lite-20150317/. W3C, 2015-03.