Schema.org
- hakukonejättien semanttinen web

Minna Tamper, Arttu Oksanen ja Eero Hyvönen
Aalto-yliopisto, Semanttisen laskennan tutkimusryhmä (SeCo), http://seco.cs.aalto.fi
Helsingin yliopisto, Digitaalisten ihmistieteiden keskus HELDIG, http://heldig.fi

Tiedon löydettävyyden parantaminen hakukoneissa on keskeinen haaste verkkoviestinnässä ja -mainonnassa. Kesäkuussa 2011 semanttisen webin maailma vavahti. Suurimmat hakukoneoperaattorit Google, Microsoft ja Yahoo! julkaisivat yhteisen Schema.org -ontologian ja sanaston, jolla voi kuvata verkkosivujen sisältöä niin, että hakukoneet löytävät ne älykkäämmin ja voivat esittää hakutuloksia havainnollisessa rakenteisessa muodossa. Maailmalla jo yli 10 miljoonaa organisaatiota hyödyntää sivuillaan Schema.org:n tarjoamia mahdollisuuksia, mutta Suomessa vain harva taho lienee tietoinen asiasta. Artikkelissa kerrotaan, miksi semanttisia merkkauksia kannattaa käyttää ja miten niitä voidaan muodostaa.

Semanttisen datan idea

Schema.org[1] perustuu semanttisen webin ideaan, jossa ihmislukijalle tarkoitettujen verkkosivujen sisällöt julkaistaan myös tietokoneiden ymmärtämässä muodossa (meta)datana, ns. semanttisena verkkona. Koko webin tasolla tästä käytetään nimitystä tiedon verkko, Web of Data, tai Giant Global Graph, GGG, erotuksena perinteiselle ihmisten World Wide Webille, WWW. Google käyttää nimitystä Knowledge Graph omasta tiedon verkostaan. Ideana on, että kun kone ymmärtää paremmin webin sisältöjä, tietoaineistojen yhdisteleminen rikkaammiksi kokonaisuuksiksi eli linkitetyksi dataksi (Linked Data) helpottuu samoin kuin tekoälyyn perustuvien verkkosovellusten kehittäminen.

Linkitetty data voidaan julkaista käytettäväksi erityisten linkitetyn datan SPARQL-palvelupisteiden ja kyselyrajapintojen kautta. Esimerkkejä tästä ovat muun muassa Wikipedioista louhittu DBpedia, Wikidata ja Suomessa Linked Data Finland -palvelun[2] kymmenet SPARQL-palvelupisteet. Toinen tapa on julkaista linkitettyä dataa verkkosivuille upotettuna hakukoneiden ja muiden sovellusten hyödynnettäväksi. Schema.org on ontologia ja tietomalli, jolla verkkosivujen julkaisijat voivat tehdä tämän mahdollisimman yksinkertaisella ja yhteentoimivalla tavalla.

[1] Määrittelyn kotisivu aineistoineen ja työkaluineen on osoitteessa http://schema.org.
[2] http://ldf.fi

Semantiikka perustuu ontologiaan

Schema.org sanasto koostuu nykyisellään 597 luokkamäärittelyistä (engl. *type*), näiden 867 ominaisuudesta (engl. *property*) sekä 114 erityisarvosta (engl. *enumeration value*) ominaisuuksille. Luokat muodostavat laajan, luokasta *Thing* alkavan luokkahierarkian, jota kautta yläluokkien ominaisuudet periytyvät alaluokille ja lopulta näiden yksilöille. Esimerkiksi luokalla *CreativeWork* on 44 alaluokkaa, kuten *Article*, *Blog*, *Comment* ja *Conversation*, joilla voi olla edelleen omia alaluokkiaan. Esimerkkinä luokan määrittelystä kuvassa 1 esitetään *CreativeWork*-luokka. Luokan merkitys kuvataan luokan määrittelyyn kuuluvien ominaisuuksien avulla. Näiden lisäksi luokka perii sen yläluokkien (tässä *Thing*) ominaisuudet. Ominaisuuksien arvoilla voi olla rajoituksia. Esimerkiksi *author*-ominaisuuden arvon on oltava yksilö luokasta *Person* tai *Organization*. Tieto verkkosivujen sisällöstä esitetään luomalla yksilöitä Schema.org-luokkahierarkian luokista ja upottamalla kuvaukset HTML-merkkausten sekaan. Näiden ja Schema.org-hierarkian avulla hakukone voi mm. löytää hakusanoilla “creative works” erilaiset luovan työn tulokset, kuten kirjat ja elokuvat, ja ominaisuuksien avulla hakutulokset voidaan esittää fiksusti jäsennettyinä “rich snippet” -tietolaatikkoina käyttäjälle.

CreativeWork

Canonical URL: http://schema.org/CreativeWork

Thing > **CreativeWork**

The most generic kind of creative work, including books, movies, photographs, software programs, etc.

Usage: Between 250,000 and 500,000 domains

[more...]

Property	Expected Type	Description
Properties from CreativeWork		
about	Thing	The subject matter of the content. Inverse property: subjectOf.
accessMode	Text	The human sensory perceptual system or cognitive faculty through which a person may process or perceive information. Expected values include: auditory, tactile, textual, visual, colorDependent, chartOnVisual, chemOnVisual, diagramOnVisual, mathOnVisual, musicOnVisual, textOnVisual.
accessModeSufficient	Text	A list of single or combined accessModes that are sufficient to understand all the intellectual content of a resource. Expected values include: auditory, tactile, textual, visual.
accessibilityAPI	Text	Indicates that the resource is compatible with the referenced accessibility API (WebSchemas wiki lists possible values).
accessibilityControl	Text	Identifies input methods that are sufficient to fully control the described resource (WebSchemas wiki lists possible values).
accessibilityFeature	Text	Content features of the resource, such as accessible media, alternatives and supported enhancements for accessibility (WebSchemas wiki lists possible values).
accessibilityHazard	Text	A characteristic of the described resource that is physiologically dangerous to some users. Related to WCAG 2.0 guideline 2.3 (WebSchemas wiki lists possible values).
accessibilitySummary	Text	A human-readable summary of specific accessibility features or deficiencies, consistent with the other accessibility metadata but expressing subtleties such as "short descriptions are present but long descriptions will be needed for non-visual users" or "short descriptions are present and no long descriptions are needed."
accountablePerson	Person	Specifies the Person that is legally accountable for the CreativeWork.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
alternativeHeadline	Text	A secondary title of the CreativeWork.
associatedMedia	MediaObject	A media object that encodes this CreativeWork. This property is a synonym for encoding.
audience	Audience	An intended audience, i.e. a group for whom something was created. Supersedes serviceAudience.
audio	AudioObject	An embedded audio object.
author	Organization or Person	The author of this content or rating. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.

Kuva 1: *CreativeWork*-luokan määrittely Schema.org-sanastossa.

Schema.org-yhteisön toimesta luokkahierarkiaa voidaan laajentaa ja tarkentaa uusille sovellusalueille määrittelemällä uusia alaluokkia ja hyväksyttämällä ne osaksi yhteistä ontologiakokonaisuutta. Esimerkiksi GoodRelations-laajennus mahdollistaa muun muassa yritysten aukioloaikojen esittämisen verkossa.

Syntaksi: formaatit

Semanttisten kuvausten tuottamiseksi verkkosivuille on olemassa useita eri formaatteja, joita käyttäen Schema.org-kuvauksia voidaan upottaa suoraan joko verkkosivun otsikko- ja metatietoihin (*head*) tai sisältöön (*body*). Formaateista keskeisimmät ovat:

- **JSON-LD**[3]**.** Merkkaustapa perustuu ohjelmistokehittäjien suosimaan JSON-formaattiin (*JavaScript Object Notation*), jota on sovellettu linkitetyn datan (LD) esittämiseen. JSON-LD-merkkauksen voi upottaa verkkosivulle <script>-tagien sisään. Google suosittelee käyttämään JSON-LD-formaattia.
- **RDFa**[4] **ja Mikrodata**[5]**.** Nämä merkkaukset lisätään sivun HTML-elementteihin erillisinä attribuutteina ja niiden arvoina.

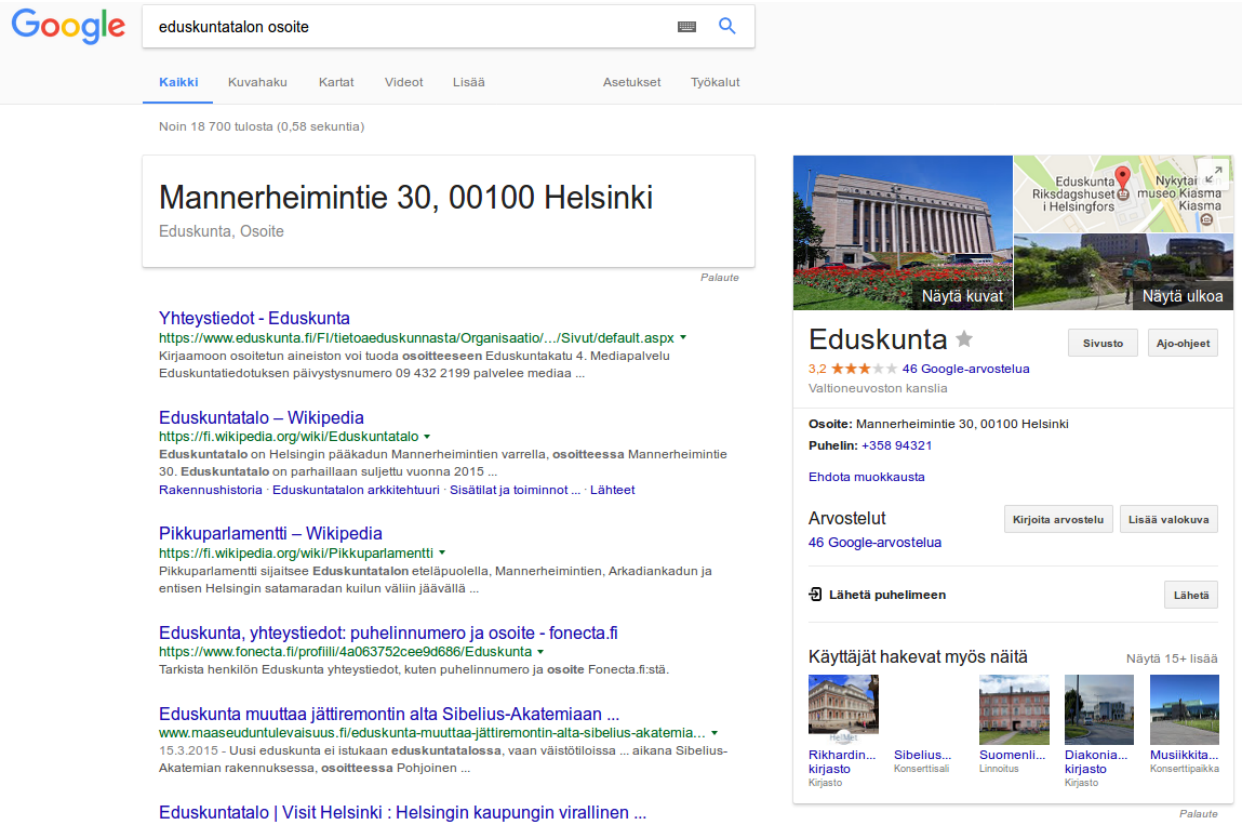
Miten hyödynnät Schema.org-merkkauksia

Parantaaksesi sivustosi näkyvyyttä hakukoneissa Schema.org-kuvausten avulla, tee seuraavat asiat:

1. **Wikipedia-artikkeli.** Laadi tarpeen mukaan sivustasi Wikipedia-artikkeli. Google käyttää Schema.org-kuvausten lisäksi myös Wikipediaa tietämysgraafinsa muodostamiseen.
2. **Wikidata-kohde.** Luo sivustasi Wikidata-kohde. Google käyttää myös Wikidatan rakenteista tietoa tietämysgraafissaan.
3. **Schema.org-kuvaus.** Lisää Schema.org-kuvaus sivullesi käyttäen jotakin edellä mainituista formaateista. Schema.org-kuvauksen voi tehdä joko käsin tai käyttämällä esimerkiksi Semanttisen laskennan tutkimusryhmän kehittämää työkalua[6]. Kuvauksen syntaksin tarkistamisessa ja testaamisessa voi käyttää apuna Googlen kehittämää jäsenneltyjen tietojen testaustyökalua[7] sekä rich-tulosten testiä[8]. Rich-tulosten testi on vielä beta-versio ja tällä hetkellä testi tukee vain Schema.org-luokkia *Recipe*, *Course*, *Movie* ja *JobPosting*.

[3] https://json-ld.org/
[4] https://rdfa.info/
[5] https://html.spec.whatwg.org/multipage/microdata.html
[6] https://semanticcomputing.github.io/json-ld-generator/
[7] https://search.google.com/structured-data/testing-tool
[8] https://search.google.com/test/rich-results

Jotta muutokset tulisivat voimaan nopeammin, voit pyytää sivuston uudelleen indeksointia hakukoneelta. Tämän jälkeen Google voi näyttää sivusi tiedot tietolaatikkona vastaavalla tavalla kuin eduskunnan tiedot kuvassa 2 oikeassa sarakkeessa.



Kuva 2: *Google-kysely, joka hakee eduskuntatalon osoitteen ja metadataan perustuvan tietolaatikon.*

Tässä artikkelissa esiteltyä tutkimusta on rahoittanut Viestintäalan tutkimussäätiö. Laajempi raportti aiheesta julkaistaan sivulla http://seco.cs.aalto.fi/publications.

Kirjallisuutta

Tom Heath ja Christian Bizer: *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool, 2011.

Eero Hyvönen: Semanttinen web, Linkitetyn avoimen datan käsikirja. Gaudeamus, 2018.