

Eero Hyvönen

Semanttinen web

Linkitetyn
avoimen
datan
käsikirja



Gaudeamus



WSOY:n kirjallisuussäätiö on tukenut teoksen kirjoittamista

Copyright © 2018 Eero Hyvönen & Gaudeamus

Gaudeamus Oy

www.gaudeamus.fi

Kansi: Pekka Krankka

KL: 61.72

UDK: 004

ISBN 978-952-495-460-0

Painopaikka: Printon Trükikoda, Tallinna 2018

Sisällys

1 Johdanto	15
I KOHTI SEMANTTISTA WEBIÄ	17
2 Webin kehityssuuntia	19
2.1 Web julkaisukanavana	19
2.2 Tiedon avoimuus – Open Data	21
2.3 Tiedon yhteisöllisyys – Web 2.0	23
2.4 Tiedon määrä – Big Data	24
2.5 Tieto palveluina – Web Services	25
2.6 Tiedon verkko – Web of Data	26
3 Tiedonhaun haasteita	32
3.1 Perinteinen tekstihaku	33
3.2 Haun haasteita	34
3.3 Selailun haasteita	38
3.4 Ratkaisuna tiedon semantiikka	39
II LINKITETTY DATA	49
4 Linkitetyn datan esittäminen	51
4.1 Tiedon esitys semanttisena verkkona	52
4.1.1 Verkko kaaviona	52
4.1.2 Verkon esittäminen kolmikkoina	52

4.1.3	Verkon looginen tulkinta	53
4.1.4	Verkko tekstinä	54
4.2	Semanttisen webin tekninen perusta	55
4.2.1	HTML-kieli	56
4.2.2	HTTP-protokolla	56
4.2.3	Yhtenäiset osoitteet ja tunnisteet: URI	57
4.2.4	Resurssien ja literaalien esittäminen	62
4.2.5	Datan linkitys tietojoukkojen välillä	64
5	RDF-verkon esittämiskielet	67
5.1	Data kolmikkoina: N-Triples	67
5.2	Yksinkertaisempaa koodia: Turtle	69
5.2.1	Nimiavaruudet	69
5.2.2	Monta samaa ominaisuutta	72
5.2.3	Eri ominaisuudet samalla kertaa	72
5.2.4	Ominaisuuden arvo omilla ominaisuuksilla	72
5.2.5	Tietotyypit	74
5.3	Verkon esittäminen RDF/XML-kielellä	76
5.4	Esimerkki tietojen yhdistämisestä	76
5.5	Muita RDF:n ominaisuuksia	81
5.5.1	Säiliöt	81
5.5.2	Kokoelmat	82
5.5.3	Reifikaatio	83
5.6	JSON-LD	84
6	Tiedon haku ja ylläpito: SPARQL	89
6.1	SPARQL-datapalvelun perustaminen	90
6.2	Peruskyselyt	92
6.2.1	Tiedon haku: SELECT	93
6.2.2	Tiedon testaaminen: ASK	98
6.2.3	Resurssin kuvaus: DESCRIBE	99
6.2.4	Uuden RDF-verkon luominen: CONSTRUCT	99
6.3	Laajennuksia peruskyselyihin	100
6.3.1	Ominaisuuspolut	100
6.3.2	Arvosijoitus	101

6.3.3	Aggregaattikyselyt	102
6.3.4	Alikyselyt	106
6.3.5	Federoitu kysely	106
6.4	Graafien hallinta ja päivitys	107
6.5	Graafin datan päivittäminen	108
7	Linkitetyn datan julkaiseminen	110
7.1	Linkitetyn datan neljä periaatetta	110
7.2	HTTP-kutsujen dereferointi	111
7.3	URI-tunnisteiden sisältöneuvottelu	113
7.4	Datajulkaisujen tähtiluokitus	115
7.5	Datan julkaiseminen WWW-sivuilla	116
III	TIETÄMYKSEN ESITTÄMINEN	121
8	Metadata ja ontologiat	123
8.1	Metadatan käsite ja muodot	124
8.2	Dublin Core -malli	126
8.3	CIDOC CRM: datan harmonisointi	129
8.4	Ontologian käsite	133
9	RDF Schema	135
9.1	Luokat ja yksilöt	135
9.2	Luokkien hierarkia	136
9.3	Ominaisuuksien hierarkia	136
9.4	Alue- ja arvorajoitteet	137
10	SKOS – Simple Knowledge Organization System	139
10.1	Käsitelmä	140
10.2	Käsitteen nimikkeet	140
10.3	Notaatiot	143
10.4	Semanttiset relaatiot	144
10.5	Ryhmittelevät termit	146
10.6	SKOS-sanastojen siltaaminen	147
10.7	Dokumentointiominaisuudet	148

10.8 Päätelysäännöt ja integriteettiehtoja	148
11 Web Ontology Language OWL	150
11.1 Johdatus logiikkaan	151
11.1.1 Logiikan idea	151
11.1.2 Lauselogiikka	152
11.1.3 Predikaattilogiikka	154
11.2 OWL-kielen syntaksit	159
11.3 Ontologiadokumentti	161
11.4 Luokkien määrittely	162
11.4.1 Samuus ja eriyys	163
11.4.2 Luokkien määrittely joukko-opillisesti	164
11.4.3 Luokkien määrittely joukko-opilla	167
11.5 Ominaisuuksien määrittely	167
11.5.1 Objektiominaisuudet	168
11.5.2 Tietotyyppiominaisuudet	170
11.5.3 Annotointiominaisuudet	171
11.6 OWL-profililit	171
11.6.1 OWL 2 EL	172
11.6.2 OWL 2 QL	172
11.6.3 OWL 2 RL	173
12 Päätelysäännöt	174
12.1 Semanttisen webin looginen tulkinta	174
12.2 Sääntöjen käyttö	177
12.2.1 Hornin logiikan suhde kuvailulogiikoihin	180
12.2.2 Suljetun maailman oletus	183
12.2.3 Yksikäsitteisten nimien oletus	184
12.2.4 Yhteenvedo	185
12.3 Käyttötapauksia säännöille	185
IV SOVELLUKSET JA INFRASTRUKTUURI	187
13 Sovellusten kehittäminen	189
13.1 Yleiset ja alakohtaiset sovellukset	190

13.2 Semanttisen portaalin osat	192
14 Portaalimalli yhteisölliseen julkaisemiseen	195
14.1 Kulttuuriaineistot verkossa	195
14.2 Hajautettu haku ja tiedon aggregointi	197
14.3 Edut käyttäjille	203
14.4 Edut julkaisijoille	203
14.5 Uusia haasteita	204
15 Sisällöntuotanto	205
15.1 Tekstin tunnistaminen (OCR)	206
15.2 Datamuunnokset ja linkitys	207
15.3 Merkitysten erottelu	209
15.4 Datan semanttinen validointi	210
16 Webin tietoinfrastruktuurit	212
16.1 Tietoinfrastruktuurin osat	213
16.2 Ontologiapalvelut	216
16.3 Työn arviointia	222
16.4 Linkitetyn datan palvelut	223
17 Sovellusesimerkki: Sotasampo	226
17.1 Tavoitteet ja käyttötapaukset	227
17.2 Aineistot ja tiedon tuottajayhteisö	228
17.3 Metadatan ja ontologiat	229
17.4 Entiteettien automaattinen linkitys	232
17.5 Portaalisovellus Sotasampo.fi	236
17.6 Datapalvelu Sotasampo	242
17.7 Uutuusarvo ja jatkokehitys	242
Liite: Esimerkki OWL-ontologiasta	245
Kirjallisuutta	251
Kirjoittaja	271