Linked Open Data: De vijf sterren van open data

Koster, L.

Published in:
Informatie Professional

De vijf sterren van open data

Linked Data zonder Open Data is slechts een theoretisch concept. Wat heb je aan een geïntegreerde informatieinfrastructuur zonder dat je de aanwezige informatie kunt gebruiken? Bij Open Data gaat het zowel om juridische toestemming voor het raadplegen en gebruiken van data, als om het daadwerkelijk mogelijk maken daarvan. De gradaties van openheid zijn door www-godfather Tim Berners-Lee uitgedrukt in Vijf Sterren, die in deze aflevering worden beschreven.

Door: **Lukas Koster**

Voor we ons verdiepen in de mate van openheid moeten we de vraag beantwoorden wat onder 'data' in dit verband wordt verstaan. De term 'data' ('gegevens') betekent informatie over fysieke en virtuele objecten. Fysieke objecten zijn bijvoorbeeld treinen, gebouwen, boeken en bankbiljetten. Virtuele objecten zijn dienstregelingen, organisaties, teksten en overheidsbudgetten. Men kan data vanuit verschillende, overlappende perspectieven karakteriseren, zoals beschrijvende data (gewicht, lengte, aantal pagina's, vervaardiger, onderwerp), gebruiksdata (aantal passagiers, uitleningen, bezoekers), statische data (gewicht, lengte van objecten), dynamische data (gewicht, lengte van levende wezens, temperatuur, bezit), intrinsieke data (gewicht, lengte, betrekking hebbend op het object zelf), relationele data (auteur, bezit, onderwerp).

Het is niet altijd mogelijk een duidelijke scheiding aan te brengen tussen fysieke en virtuele objecten. Is een boek alleen het fysieke exemplaar of ook de tekstuele inhoud? Een tekst bestaat ook uit informatie en data, maar teksten en andere objecten waarop auteursrecht van toepassing is, vormen een apart domein. Men spreekt dan niet over 'open data', maar over 'open access'. Hierbij dienen data om toegang tot of informatie over die objecten te verkrijgen. Men spreekt dan gewoonlijk van 'metadata'. In het andere geval bestaan virtuele objecten juist uit data. Een dienstregeling bijvoorbeeld is een samenstel van data over routes, begin- en eindpunten, halteplaatsen, vertrek- en aankomsttijden, soort vervoermiddel, et cetera.

Volledige openheid

De juridische en de praktische kant van Open Data worden weerspiegeld in de Vijf Sterren van Open Data, in 2010 geformuleerd door Tim Berners-Lee, de bedenker van het World Wide Web HTTP-protocol en tevens de opsteller van de Vier Regels van Linked Data (http://www.w3.org/DesignIssues/LinkedData.html). De Vier Regels hebben betrekking op da basisstandaarden HTTP URI's, RDF, SPARQL en het opnemen van externe links. Deze komen weer terug in de vierde en vijfde ster. De vijf sterren van open data beschrijven de stappen om tot volledige openheid te komen. Elke volgende ster omvat de vereisten van de voorgaande sterren. >

Beschikbaar op het web (in welke vorm dan ook), maar met een open licentie

Fysieke databestanden (zoals kaartenbakken, kadastermappen of archiefstukken) kunnen openbaar raadpleegbaar zijn, maar ze zijn niet altijd en overal voor iedereen beschikbaar. Als pagina, object of bestand op het web zijn ze dat wel. Maar om de eerste ster te verdienen moet deze informatie op zijn minst gratis toegankelijk zijn, door middel van een open licentie. Op data op zichzelf rust geen wettelijke bescherming, maar op dataverzamelingen is in het algemeen een vorm van auteursrecht van toepassing als er een substantiële inspanning gedaan is voor het samenstellen ervan.

Er zijn diverse soorten open licenties voor raadplegen en hergebruik van informatie onder verschillende voorwaarden. Veel gebruikt zijn de Creative Commons-licenties (http://creativecommons.org/licenses), oorspronkelijk gericht op objecten waarop auteursrecht van toepassing is. Met de introductie van versie 4.0 zijn de CC-licenties ook bruikbaar voor data en databases (http://creativecommons.org/Version4). CC kent vier categorieën die onderling gecombineerd kunnen worden, te weten: 'Attribution' (BY): naams- of bronvermelding verplicht; 'Share Alike' (SA): mag onder dezelfde voorwaarden verspreid of gebruikt worden; 'No Derivatives' (ND): mag alleen in zijn geheel zonder wijzigingen verspreid of gebruikt worden; 'Non Commercial' (NC): zonder commerciële doeleinden. De CC BY-NC-ND is de meest restrictieve licentie, CC BY de meest vrije. Daarnaast is er de CC0 (CC Zero) publieke domeinverklaring die een stapje verder gaat. Hiermee is alles zonder bronvermelding toegestaan.

Het gebruik van 'Non Commercial' is omstreden, omdat het begrip niet eenduidig gedefinieerd is. Het toepassen van een NC-licentie betekent niet alleen dat er geen geld met de hergebruikte informatie mag worden verdiend, maar ook dat een commerciële organisatie ze niet in gratis producten en diensten mag gebruiken. 'Non Commercial' impliceert een restrictie en is dus de facto niet open.

Voor data en databases worden ook andere licenties gebruikt, met name de Open Data Commons ODC (http://opendatacommons.org): 'Public Domain Dedication and License' (PDDL), 'Open Data Commons Attribution License' (ODC-BY) en 'Open Database License-Attribution and Share Alike License' (ODC-ODbL).

Belangrijke vraag is of bronvermelding op het niveau van de aggregatie of van de individuele data vereist is. In het laatste geval is het namelijk veel moeilijker te implementeren en impliciet minder open.

Beschikbaar als machine-leesbare gestructureerde data (bijvoorbeeld Excel in plaats van een beeldscan van een tabel)

Een ingescande afbeelding van een cataloguskaart, een PDF van een kadasterpagina, een foto van een archiefstuk en zelfs een online bibliotheekcatalogus met een open licentie kunnen wel als Open Data worden gekenschetst, maar deze objecten zijn alleen te lezen en te verwerken door middel van handmatige menselijke tussenkomst. Eén ster dus. Als dezelfde informatie in gestructureerde vorm beschikbaar en door software automatisch of semi-automatisch te verwerken is, zijn we een stapje verder op weg naar herbruikbaarheid. Een ster erbij! Impliciet wordt aangenomen dat een beschrijving van de data en de structuur beschikbaar is. Zonder zo'n beschrijving kan de informatie niet geïnterpreteerd worden. Berners-Lee heeft dat later in een extra notitie onder de noemer 'metadata' toegevoegd.

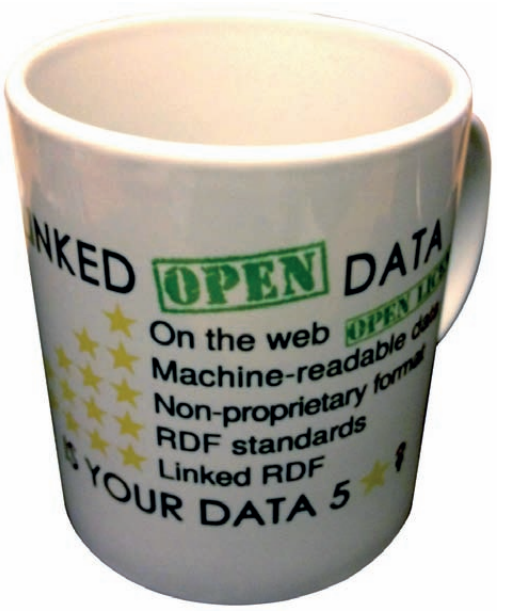
'Om de eerste ster te verdienen moet de informatie op zijn minst gratis toegankelijk zijn, door middel van een open licentie'

Als (2), maar in een niet-systeemgebonden vorm (bijvoorbeeld CSV in plaats van Excel)

Als voor het raadplegen en verwerken van gestructureerde data met een open licentie speciale commerciële of niet meer verkrijgbare software nodig is, dan zijn die data de facto niet toegankelijk voor mensen die niet over die software kunnen beschikken. Het hier genoemde Excel is niet meer zo'n goed voorbeeld, omdat met de meeste databasesoftware Excel-bestanden geïmporteerd kunnen worden. Het gaat erom dat de data geleverd worden in een algemeen bruikbaar format. Drie sterren! Voor de normale praktijk is dit genoeg: data beschikbaar op het web onder een open licentie in een open machine-leesbaar format. Het is niet helemaal duidelijk of hier alleen gedoeld wordt op het downloaden van volledige databestanden ineens, of ook op deelselecties en individuele records. Verder wordt niet gespecificeerd of ook 'on the fly'-gebruik bedoeld wordt. Dit laatste wordt doorgaans mogelijk gemaakt door middel van API's (application programming interfaces), aparte webtoegangen waarmee externe systemen data in gestructureerde vorm voor onmiddellijke verwerking kunnen opvragen.

Al het bovenstaande, plus het gebruik van open standaarden van W3C (RDF en SPARQL) om dingen te identificeren, zodat mensen naar je spullen kunnen verwijzen

Met de vierde ster wordt de stap gezet naar Linked Data, waardoor we uitkomen bij Linked Open Data. RDF is hierin het universele, open, uitbreidbare, niet-systeemgebonden, gestructureerde dataformat met URI's als unieke sleutels voor objecten en data op het web. SPARQL is de vraagtaal voor databases in RDF-format, vergelijkbaar met SQL voor relationele databases. Een SPARQL Endpoint is de systeemonafhankelijke versie van een API. Waar je via een API communiceert met een systeem om data op te vragen, met alleen de binnen dat systeem beschikbare opties, heb je via SPARQL rechtstreeks toegang tot de data.



Al het bovenstaande, plus: link je data aan andermans data om context te verschaffen

Met het publiceren van data onder een open licentie als RDF met directe toegang via URI's op het web zijn alle toegangsdeuren wagenwijd opengezet. Maar er is nog geen uitgang. Door het vervangen van tekstwaarden, zoals auteursnaam, door links naar geautoriseerde thesauri en het toevoegen van links naar objecten in andere databronnen, zoals DBPedia, worden bruggen geslagen en nieuwe wegen geopend naar gerelateerde informatie die elders beschikbaar is. Met de vijfde ster staan alle ramen open.

Lukas Koster is Coördinator Bibliotheeksystemen bij de Bibliotheek van de Universiteit van Amsterdam.