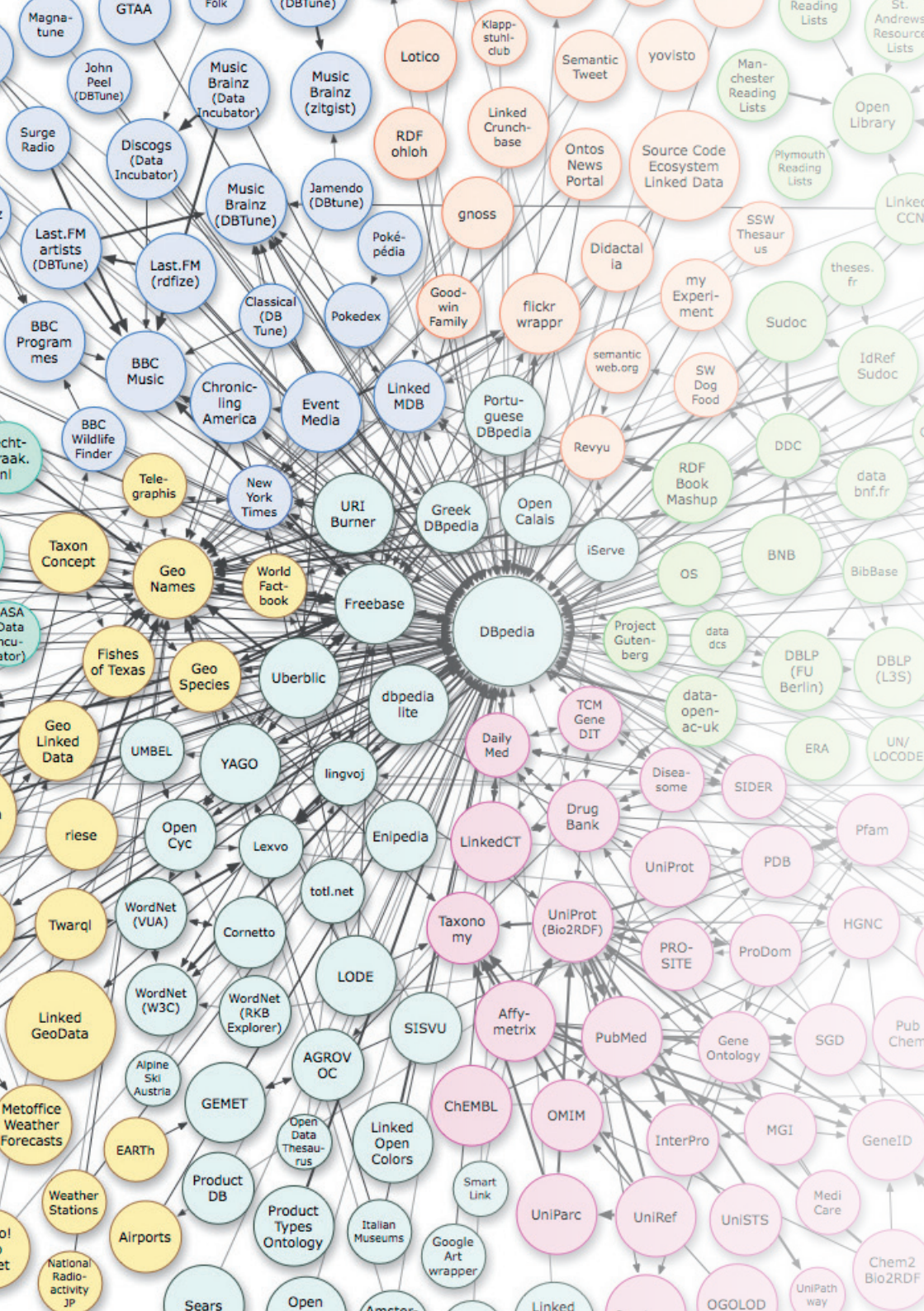
# Data, Data, Data: Big, Linked & Open


Magna-tune
GTAA
Folk
(DBTune)
John Peel (DBTune)
Music Brainz (Data Incubator)
Music Brainz (zitgist)
Lotico
Klapp-stuhl-club
Semantic Tweet
yovisto
Reading Lists
St. Andrews Resource Lists
Man-chester Reading Lists
Open Library
Surge Radio
Discogs (Data Incubator)
RDF ohloh
Linked Crunch-base
Ontos News Portal
Source Code Ecosystem Linked Data
Plymouth Reading Lists
Linked CCN
Music Brainz (DBTune)
Jamendo (DBtune)
gnoss
SSW Thesaur us
Last.FM artists (DBTune)
Last.FM (rdfize)
Poké-pédia
Didactal ia
theses. fr
Classical (DB Tune)
Pokedex
Good-win Family
flickr wrappr
my Experi-ment
Sudoc
BBC Programmes
BBC Music
Chronic-ling America
Event Media
Linked MDB
Portu-guese DBpedia
semantic web.org
SW Dog Food
IdRef Sudoc
BBC Wildlife Finder
Revyu
DDC
Tele-graphis
New York Times
URI Burner
Greek DBpedia
Open Calais
RDF Book Mashup
data bnf.fr
Taxon Concept
Geo Names
World Fact-book
Freebase
DBpedia
iServe
OS
BNB
BibBase
Project Guten-berg
data dcs
Fishes of Texas
Geo Species
Uberblic
dbpedia lite
DBLP (FU Berlin)
DBLP (L3S)
Geo Linked Data
UMBEL
YAGO
lingvoj
TCM Gene DIT
Daily Med
data-open-ac-uk
ERA
UN/ LOCODE
Disea-some
SIDER
riese
Open Cyc
Lexvo
Enipedia
Drug Bank
Pfam
LinkedCT
UniProt
PDB
totl.net
WordNet (VUA)
Twarql
Cornetto
Taxono my
UniProt (Bio2RDF)
PRO-SITE
ProDom
HGNC
WordNet (W3C)
WordNet (RKB Explorer)
LODE
Affy-metrix
PubMed
Gene Ontology
SGD
Linked GeoData
SISVU
AGROV OC
Alpine Ski Austria
GEMET
ChEMBL
OMIM
InterPro
MGI
GeneID
Metoffice Weather Forecasts
EARTh
Open Data Thesau-rus
Linked Open Colors
Smart Link
Medi Care
Weather Stations
Product DB
UniParc
UniRef
UniSTS
Airports
Product Types Ontology
Italian Museums
Google Art wrapper
Chem2 Bio2RDF
National Radio-activity JP
Sears
Open
OGOLOD
UniPath way

# Data, Data, Data: Big, Linked & Open

Auteurs

**Erwin Folmer (Geonovum)**
**Dennis Krukkert (TNO)**
**Silja Eckartz (TNO)**

**De gehele business en IT-wereld praat op dit moment over Big Data, een trend die medio 2013 Cloud Computing is gepasseerd (op basis van Google Trends). Ook beleidsmakers houden zich actief bezig met Big Data.**

Neelie Kroes, vice-president van de Europese Commissie, spreekt over de 'Big Data Revolution' en de verandering naar een 'Data-driven Economy', waarbij data als de nieuwe olie worden beschouwd die tientallen miljarden euro's waard zijn. [1]

Deze Big Data Revolutie wordt ook uitmuntend beschreven in het boek 'De Big Data Revolutie', waarin big data wordt beschreven als bron van economische waarde en innovatie: 'Gegevens kunnen op een slimme manier worden hergebruikt en zo uitgroeien tot een rijke bron van innovaties en nieuwe diensten'. Een aspect daarbij is de gigantische toename van data en de noodzakelijke machines, maar de echte revolutie zit in de gegevens zelf en de gebruikswijze. [2]

De bedrijvigheid en innovatieve aspecten worden breed gedragen, en in een overheidscontext vaak ook gerelateerd aan Open Data in het publieke domein. Zoals Kroes het heeft over 'Unlocking this gold mine' en 'Opening up public data means opening up business opportunities'. Maar ondanks de stevige uitspraken van Kroes is Europa geen koploper.

In mei 2013 heeft Obama een nieuwe wet voor Open Data ondertekend: 'And today I'm announcing that we're making even more government data available, and we're making it easier for people to find and to use. And that's going to help launch more start-ups. It's going to help launch more businesses... It's going to help more entrepreneurs come up with products and services that we haven't even imagined yet. This kind of innovation and ingenuity has the potential to transform the way we do almost everything.' Daarbij gaat Obama verder dan alleen het open publiceren: een open licentie, machine readable, en in een open formaat. Daarnaast zal de data gecombineerd

en gerelateerd moeten worden voor toepassingen. **[3]**

Ook het OECD heeft een rapport **[4]** gepubliceerd over Big Data, of eigenlijk de economische mogelijkheden van de nieuwe data society. Daarin wordt een vijftal toepassingen of trends beschreven:

- Data-driven R&D: verbetering in onderzoek door data toepassing
- Data(-intensive) products: het ontwikkelen van nieuwe producten/diensten waarin data het product is (of component)
- Data-driven processes: optimaliseren van productieprocessen
- Data-driven marketing: meer gerichte advertenties en gepersonaliseerde aanbevelingen
- Data-driven organisation: ontwikkelen van nieuwe benaderingen voor het management in organisaties

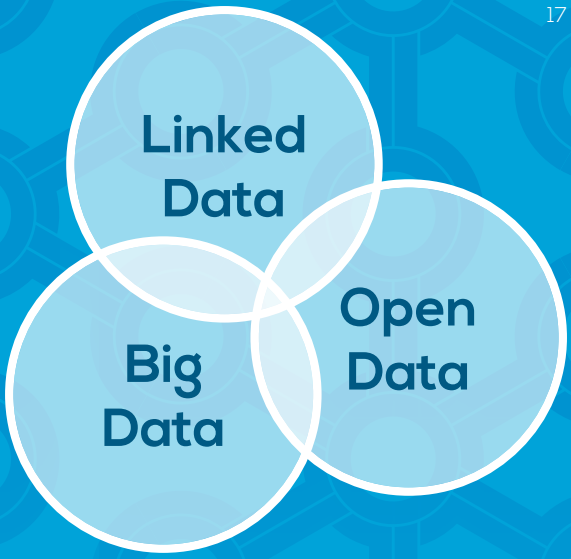
Maar is dit nu Big Data? De terminologie in ICT is al jaren een hot issue; termen en trends met spannende afkortingen zoals een SOA, volgen elkaar in rap tempo op. Zoals gezegd in 2013 is Big Data booming. Linked Open Data is een term die al een lange geschiedenis heeft en al jaren relatief stabiel is, maar het is sterk gerelateerd aan de trend Big Data, en uiteraard ook aan Open Data.

In het kort samengevat gaat Big Data over de slimme toepassing van data in alle soorten en maten die tot op heden nog niet mogelijk waren. Eigenlijk is de term 'big' dus misleidend. We hadden het beter gewoon 'Data, Data, Data' kunnen noemen zoals in de Roadmap ICT; dat dekt beter de lading. **[5]**

Regelmatig worden ook andere termen geïntroduceerd om de lading beter te dekken. Zoals Small Data **[6]**, Long Data **[7]** of Smart Data **[8]**, en voor specifieke toepassings-domeinen zoals Linked Enterprise Data. Maar het gaat om data, data, data.

## De viewpoints op 'Data, Data, Data'

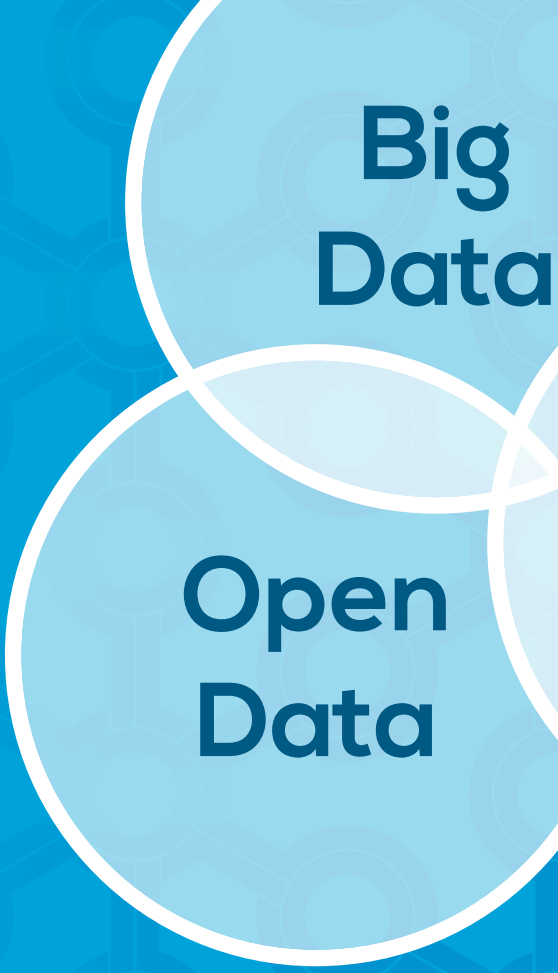
Er zijn vele 'viewpoints' mogelijk op 'data', maar de drie belangrijkste zijn toch wel Big, Open en Linked; deze overlappen elkaar ook. Op de volgende pagina's worden de viewpoints nader beschreven. De drie viewpoints hebben overlap, en waar ze elkaar overlappen wordt het vaak interessant. Vaak gaan initiatieven van Linked Data over vrij beschikbare (open) data, en spelen issues gerelateerd aan het Big viewpoint ook een rol.




## Big Data

De benaming 'Big' Data view mag enigszins verwarrend overkomen, want in de vorige paragraaf lieten we zien dat de hype Big Data, niet heel specifiek over 'Big' gaat. Echter binnen de ruime hype term (die wij liever data, data, data noemen) is er zeker wel een onderwerp dat specifiek over 'Big' gaat. Meestal worden dan de drie V's (volume, velocity, en variety) gebruikt om aan te tonen dat het 'Big' is, ook al zijn er geen eisen aan bijvoorbeeld het Volume gedefinieerd voordat we over 'Big' zouden moeten praten. Maar als je aan een informaticus vraagt wat Big Data is dan komt hij met een krapper, technischer antwoord: Big data bestaat uit datasets die te groot zijn om met reguliere databasemanagementsystemen te onderhouden. Denk hier bijvoorbeeld aan sensor data of data streams.

Hier wordt ook de term NoSQL voor gebruikt om uit te drukken dat het om ongestructureerde data in schemaloze databases gaat. Voor een informaticus spelen dan ook data mining en data analytics een heel belangrijke rol voor big data. Verschillende open source softwaremodellen ondersteunen data-intensieve gedistribueerde applicaties. Een voorbeeld is MapReduce, een raamwerk voor het verwerken van problemen over enorme datasets met behulp van een groot aantal computers (nodes), ook cluster of grid genoemd. Een populaire vrije implementatie is Apache Hadoop. Hierbij gaat het ook over process hosting: waar worden de data-analyseprocessen uitgevoerd (bijvoorbeeld Amazon EC2)? En waar staat de data (cloud computing, distributed hosting (bijvoorbeeld Amazon S3)?




## Open Data

'Open' Data zijn datasets die met een open licentie beschikbaar worden gesteld zodat toegang en hergebruik zonder beperkingen mogelijk is. De voorwaarden waaronder de data beschikbaar zijn, worden beschreven in licenties en gebruiksvoorwaarden. Het idee van open data is om de beperkingen in hergebruik tot een minimum te limiteren. Hierdoor wordt het delen en hergebruik van data bevordert. Vaak wordt data van de overheid op deze manier gedeeld om transparantie te vergroten en economische activiteit te bevorderen. Verder geeft open data derde partijen de kans om met data te innoveren.

Veelal wordt ook onder open data verstaan dat data (indien relevant) 'machine readable' moet zijn en in een open formaat beschikbaar moet worden gesteld. In principe kunnen tekstdocumenten ook geprint worden, gescand worden en in jpg als open data beschikbaar worden gesteld. Het doel van open data, hergebruik, wordt dan wel lastig. Hetzelfde geldt voor data die beschikbaar worden gesteld in een gesloten formaat dat alleen met dure software is in te lezen; ook dat belemmert het hergebruik. Vandaar: open licentie, machine readable, en open formaat. Tot slot wordt regelmatig vergeten dat onderhoud, kwaliteit en governance op de open data ook belangrijke aspecten zijn. Datasets kunnen snel verouderen, dus alleen eenmalig op Internet zetten is niet voldoende. Wat is de kwaliteit, en wie bepaalt welke correcties wanneer doorgevoerd gaan worden?

# Linked Data

## Linked Data

'Linked' Data is een essentieel onderdeel van het Semantische Web. Door Tim Berners-Lee in 2006 al beschreven als een component van 'Web 3.0', een trend waarbij Internettoepassingen goed op elkaar afgestemd en soms zelf geïntegreerd kunnen worden. De ontwikkeling van het Internet wordt dan als volgt beschreven: van het web van documenten (Web 1.0) via Web 2.0 waar het Internet als interactief communicatiemedium beschouwd wordt en gebruikers informatie kunnen uploaden naar Web 3.0: het web van Linked Data, waar Internettoepassingen en data bijvoorbeeld via webservices aan elkaar gelinkt kunnen worden. Deze slimme toepassingen kunnen dan de links volgen tussen datasets. Dit is de basis van het semantic web.

Linked Data, Semantic Web, Web 3.0 worden weleens als synoniemen gebruikt, alhoewel specifiek Linked Data gebruikt wordt als term voor een methode voor het publiceren van data in een structuur zodat het linkbaar wordt en daarmee bruikbaar.

## Linked Open Data

Linked Open Data is een praktische manier om een bijdrage te leveren aan het Semantische Web. Semantisch wil zoveel zeggen als 'de betekenis lerend'. Het semantisch web zou je grofweg kunnen definiëren als een web van verbanden. Verbanden gelegd tussen informatie op het internet, waardoor nieuwe inzichten kunnen ontstaan. Informatie gepubliceerd als Linked Open Data stimuleert het hergebruik van je data, omdat je zelf zoveel mogelijk verwijzingen aanbrengt naar kennisbronnen elders en omdat anderen gemakkelijk naar jouw informatie kunnen verwijzen.

Als je in een willekeurige webbrowser naar 's Gravenhage zoekt, vind je geen resultaten waarin 'Den Haag' voorkomt, terwijl beide woorden naar dezelfde stad verwijzen. Dat komt doordat webdocumenten met elkaar verlinkt zijn, maar de inhoud zelf niet. Een zoekmachine kan zo alleen maar op woorden zoeken. Linked Data biedt een oplossing voor dit probleem door woorden als concepten uniek te maken en te beschrijven in één of liefst meerdere subject – predicaat – objectrelaties. Een stad wordt daarmee een concept en kan meerdere attributen krijgen, waarvan elk attribuut ook weer een eigen concept is. Zowel subject, als predicaat en object zijn dus op zichzelf weer unieke concepten. Elk concept wint aan betekenis naarmate er meer beschrijvingen aan gelinkt worden. Op deze manier wordt de inhoud

van webdocumenten betekenisvol en worden zoekresultaten nauwkeuriger. Uiteindelijk bereik je zo taalonafhankelijkheid omdat het dus niet meer uitmaakt of je naar 'The Hague' of 'Den Haag' zoekt.

Zoals gezegd staat 'Linked Data' voor het idee dat informatie (resources) op het web aan elkaar gelinkt wordt zodat je informatie maar één keer hoeft te beschrijven, en er vervolgens vanuit verschillende bronnen naar kunt verwijzen. Open data gaat over het ontsluiten van data op een zodanige manier dat eenieder die data kan gebruiken zonder vast te zitten aan licenties die het gebruiksrecht beperken. Als we beide combineren dan komen we op Linked Open Data, oftewel: data die via het web op een open manier aangeboden wordt, en die onderling verbonden is.

## Het belang van Linked Data

Om nauwkeurig te kunnen zoeken in de enorme hoeveelheid informatie op het web is zoeken op basis van betekenis onontkoombaar. Het semantisch web is een verzamelnaam voor technieken die computers in staat stellen de betekenis van de informatie op het web te begrijpen zónder menselijke tussenkomst. Een complicerende factor hierbij is dat de betekenis van mensen, dingen, gebeurtenissen etc. niet constant is, maar kan variëren. Zo heeft de koningin van Nederland een wisselende betekenis die o.a. afhankelijk is van de tijd: in 2010 verwijst het begrip naar Beatrix, maar in 1970 was het Juliana, en in 2013 naar Willem-Alexander. Menselijke informatieverwerkers zijn gewend om contextuele factoren, zoals tijd, mee te nemen bij het toekennen

van betekenis. Voor machines geldt dit niet. Om computers toch in staat te stellen de juiste betekenis toe te kennen, is het aanbieden van relevante context van groot belang. Linked Data is een techniek om machine-leesbare context te genereren.

We sluiten deze paragraaf af met een quote uit het OECD-rapport: 'The linking and use of data across sectors drive innovation, socioeconomic development and growth.' **[4]**

## Linked Open Data concepten

Om data op een betekenisvolle manier aan elkaar te linken wordt gebruik gemaakt van 'semantic web' technologie. Dit is een verzamelnaam voor verschillende standaarden en technologieën die gebruikt kunnen worden om te komen tot een web van 'Linked Open Data' die op een betekenisvolle manier gebruikt (en door computers geïnterpreteerd) kan worden.

De basis voor het semantic web is RDF (Resource Description Framework). Een resource kan van alles zijn: een persoon (bijvoorbeeld Erwin, een auteur van dit hoofdstuk), een organisatie (Geonovum, een werkgever van Erwin), een boek (bijvoorbeeld dit boek), een artikel, etc. Elke resource heeft een unieke URI in de vorm van een URL. Op de locatie van deze URL wordt een representatie van de resource aangeboden.

Het linken van data gaat via zogenaamde 'triples'. Een triple is een getypeerde relatie tussen twee resources, en bestaat uit een subject, een predicate en een object. Hiermee kun je bijvoorbeeld uitdrukken Erwin (subject) werkt voor (predicate) Geonovum (object). Erwin en Geonovum zijn beide resources waarnaar verwezen wordt middels een URI. Van het boek Pilot Linked Open Data kan ook een resource gemaakt worden. Het is nu heel eenvoudig om aan te geven wie de auteur van dit boek is door een triple toe te voegen: Pilot Linked Open Data (subject) is geschreven door (predicate) Erwin (object). Omdat elke keer verwezen wordt naar dezelfde resource hoeft deze maar één keer beschreven te worden.

Het hergebruik van data wordt pas echt waardevol door niet alleen data te linken, maar door ook betekenis toe te voegen aan de link. Oftewel: door expliciet vast te leggen wat de relatie (de 'predicate') 'werkt voor' betekent. Ook hiervoor wordt de hetzelfde mechanisme gebruikt: de relatie wordt beschreven in RDF-formaat, en ontsloten via een URI.

In aanvulling op RDF kan RDFS gebruikt worden. Met behulp van RDFS kunnen 'klassen' van resources aangemaakt worden, en tevens beperkingen gelegd worden op de verschillende relaties die mogelijk zijn tussen instanties van deze klassen. Hiermee zou je bijvoorbeeld vast kunnen leggen dat de relatie (lees: predicate) 'is getrouwd met' alleen gelegd kan worden tussen twee resources van het type 'persoon'.

Gelukkig hoeven we het wiel niet opnieuw uit te vinden. Er zijn al veel standaard definities beschikbaar van resource- en relatietypes. Deze zijn vastgelegd in zogenaamde ontologieën. Een ontologie is, eenvoudig gesteld, een neerslag van concepten en hun onderlinge verbanden die door een groep betrokkenen gezien wordt als representatie van hun gedeelde werkelijkheid. Afhankelijk van de toepassing kan een ontologie daarmee de vorm krijgen van bijvoorbeeld een definitielijst, een RDF-graph, een UML-model, een taxonomie, een zuiver logisch model en vele andere. Voorbeelden van veelgebruikte ontologieën zijn SKOS (Simple Knowledge Organization System) voor het opstellen van vocabulaires en thesauri, en FOAF (Friend-of-a-friend). SKOS bevat zaken als 'broader', 'narrower' en 'alternative label'. FOAF bevat zaken als 'Person' en 'email'.

RDF beschrijft het algemene principe achter triples, en schrijft bijvoorbeeld het gebruik van URI's voor. Het technische formaat (de syntax) waarin triples worden vastgelegd en uitgewisseld wordt echter vrijgelaten. Hiervoor zijn verschillende technische formaten beschikbaar zoals RDF/XML, Turtle, N3 en JSON. Met elk van deze formaten kunnen triples vastgelegd worden. De formaten verschillen enerzijds in syntax, en anderzijds in een aantal toevoegingen bovenop RDF. Triples kunnen ook opgeslagen worden in een database, ook wel 'triple store' genoemd. Om een triple store te bevragen is een speciale taal ontwikkeld (analoog aan SQL voor relationele database): SPARQL (SPARQL Protocol and RDF Query Language).

# De vier principes van Linked Open Data

**Linked Data moet aan vier principes (vrij naar Tim Berners Lee) voldoen [11]. Deze principes zijn:**

### Principe 1
Gebruik URI's om dingen te identificeren.

### Principe 2
Gebruik HTTP-URI's zodat er naar deze dingen kan worden verwezen en dat ze kunnen worden opgezocht door zowel mensen als machines.

### Principe 3
Leg de informatie over het concept vast in een 'triple' (subject-predicaat-objectrelatie), leg die triple vast en maak het beschikbaar op basis van standaarden zoals RDF en SPARQL.

### Principe 4
Neem links naar andere, gerelateerde, open-dataconcepten op in de beschrijving om het ontdekken van gerelateerde informatie op het web te verbeteren.

Een URI betekent in feite het toekennen van een unieke string om data-objecten uniek identificeerbaar te maken. Door middel van HTTP-URI's wordt er verwezen naar een unieke plek op het internet waardoor informatie over het object vindbaar wordt. Linked Open Data zijn datasets in RDF-formaat, met URI's en met links tussen de datasets.

# De weg naar Linked Open Data: het vijfsterrenmodel

Organisaties zullen vaak niet direct met Linked Open Data beginnen. Vaak worden er verschillende stappen gezet voordat men kan spreken over Linked Open Data. De eerste stappen betreffen veelal puur het open maken van Data. Dit heeft Tim Berners-Lee verwerkt in het vijf-sterrenmodel van Linked Open Data [12]. De vijf sterren in dit model zijn hieronder beschreven.

De eerste drie sterren betreffen open data, en pas bij de laatste twee sterren spreken we van Linked Open Data. In 2013 maken leiders zoals Kroes en Obama duidelijk een beleid gericht op drie sterren, terwijl in het verleden één ster veelal voldoende was. Als die lijn zich doorzet, zitten we binnen een paar jaar op vijf sterren (overheids)data!

De informatie is beschikbaar op het internet, in welk formaat dan ook.

De informatie is online beschikbaar in een gestructureerd formaat, dat geschikt is voor automatisch hergebruik (zoals Excel in plaats van een plaatje van een tabel).

De informatie is online beschikbaar in een open bestandsformaat (zoals CSV in plaats van Excel).

Al het bovenstaande, en bovendien wordt gebruikgemaakt van de open standaarden Resource Description Framework (RDF) en SPARQL, zodat anderen makkelijk naar de dataobjecten kunnen verwijzen.

Al het bovenstaande, en bovendien wordt er naar data van anderen verwezen voor meer context van de data.

## Referenties


**[1]** Statements Kroes:
http://bit.ly/19bpC0q

**[2]** Boek Big Data: Viktor Mayer-Schönberger & Kenneth Cukier (2013), De Big Data Revolutie, Hoe de data-explosie al onze vragen gaat beantwoorden, Maven Publishing.

**[3]** Statements Obama:
http://bit.ly/197zCG2

http://1.usa.gov/15NKxby

**[4]** OECD rapport: OECD (2013), 'Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by 'Big Data'', OECD Digital Economy Papers, No. 222, OECD Publishing.
http://bit.ly/14DIB2E

**[5]** Roadmap ICT:
http://bit.ly/ShQGEd

**[6]** Small Data:
http://bit.ly/17Q3Z88

**[7]** Long Data:
http://bit.ly/Wd9FCK

**[8]** Smart Data:
http://bit.ly/ZUgNoV

**[9]** Boek Linked Data: Tom Heath & Christian Bizer (2011), Linked Data, Evolving the Web into a Global Data Space, Morgan & Claypool Publishers.
http://bit.ly/13Qsz2W

**[10]** Boek Linked Open Data: Florian Bauer & Martin Kaltenböck (Semantic Web Company) Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers. De eerste drie hoofdstukken uit dit boek zijn opgenomen in deel 2.
Het volledige boek is beschikbaar als PDF op http://bit.ly/AC4EIZ

**[11]** Vier principes van Linked Open Data: http://bit.ly/FQVNV

**[12]** Het vijfsterren-model: http://www.5stardata.info