Uw data op het Web van Data

Een beknopt plan van aanpak voor erfgoedinstellingen

Ivo Zandhuis

In opdracht van het Provinciaal Historisch Centrum

8 maart 2011

ERFGOEDHUIS·ZH

Over het Provinciaal Historisch Centrum

Het Provinciaal Historisch Centrum (PHC) van het Erfgoedhuis Zuid-Holland zet zich in om de geschiedenis van Zuid-Holland en de daarmee verbonden erfgoedcollecties onder de aandacht te brengen van een breed publiek. Het werk van Erfgoedhuis Zuid-Holland wordt in belangrijke mate mogelijk gemaakt door financiële steun van de provincie Zuid-Holland.

Aanleiding voor dit plan van aanpak

In het kader van de pilot 'Open Bronnen' in 2010 heeft Frans van der Horst (toen student aan de Hogeschool van Amsterdam) bij het PHC zijn scriptieonderzoek uitgevoerd over het semantisch web. Hij heeft beschreven wat het semantisch web kan betekenen voor een webportal in het algemeen en voor de website www.geschiedenisvanzuidholland.nl in het bijzonder. Dit plan van aanpak is een vervolg op zijn conclusies en aanbevelingen.

1. Inleiding

1.1. Aanleiding

Al vele jaren wordt gesproken over de ontwikkeling van het Semantisch Web: het lijkt een toekomst die maar niet lijkt te komen. Toch is er met de introductie van de term Linked Open Data een beweging op gang gekomen waarin veel data nu werkelijk in deze vorm op het Semantisch Web gepubliceerd worden.

1.2. Belang

Erfgoedinstellingen beschikken over grote hoeveelheden data waarvan je kunt betogen dat deze zonder juridische of technische drempels beschikbaar zouden moeten zijn voor iedereen die daar gebruik van wil maken. De technische drempels worden door de introductie van semantisch web technologie voor een belangrijk deel geslecht.

1.3. Centrale vraag

Dit leidt als vanzelf tot de vraag: hoe kunnen erfgoedinstellingen hun data beschikbaar stellen met technologie van het semantisch web?

1.4. Aanpak

In samenwerking met het Provinciaal Historisch Centrum van Erfgoedhuis Zuid-Holland is uitgewerkt hoe de publicatie van Open Data op geschiedenisvanzuidholland.nl kan worden gerealiseerd. Dit heeft geresulteerd in een realistisch stappenplan dat ook kan worden toegepast bij andere instellingen.

2. Het Web van Data

2.1. Wat is er mis met het huidige web?

Data is nu alleen geschikt voor mensen
Het World Wide Web dat we kennen is gericht op het publiceren van documenten in de vorm van webpagina's. Soms worden deze webpagina's gemaakt op basis van een zoekvraag die door een gebruiker in een zoekformulier is ingevuld. De data in de database, waarop deze webpagina's zijn gebaseerd, blijven verborgen. Het zoekformulier beperkt de gebruiker in het stellen van een zoekvraag en beperkt andere zoekmachines bij het indexeren van de data[1].

Data is nu verschraald om geschikt te maken voor portals
De samenwerking van verschillende erfgoedinstellingen heeft ertoe geleid dat er websites zijn ontwikkeld zoals de Zuid-Hollandse erfgoedportal www.geschiedenisvanzuidholland.nl, waarin het datamodel van de portal is verschraald tot een grootste gemene deler, waarin alle data kan worden opgenomen. Dit is nodig voor het ontwikkelen van een gebruiksvriendelijk interface. Het verschralen van het datamodel leidt tot beperking van her-bruikbaarheid (duurzaamheid) van data: indien een andere meer specifieke interface moet worden gemaakt, moet een nieuwe vertaling worden gemaakt van het bron-systeem naar de portal.

Data bevat nu minder dan zou kunnen
Als er voldoende relaties worden gelegd tussen collectieonderdelen, kan steeds meer context worden gegeven over de documenten en informatie die gepresenteerd wordt. Op het huidige web zijn deze relaties moeilijk op een duurzame manier te leggen en te gebruiken, omdat de (technische) mogelijkheid ontbreekt om losse data-onderdelen naar elkaar te laten verwijzen.

2.2. Hoe kan dat dan beter?

Op het Semantisch Web publiceert u uw data door de data die nu in gestructureerde vorm in databases en/of XML-bestanden zijn opgeslagen op een specifieke manier te exporteren. De specifieke manier van exporteren levert databestanden op die volgens het semantisch web zijn gecodeerd. Vervolgens plaatst u deze databestanden online. Doordat de publicatie op het internet plaatsvindt, kan elke computer deze data en haar structuur gebruiken en kan iedereen zijn eigen zoekvraag en analyse op de data loslaten. Hierbij wordt de gebruiker en de zoekmachine niet beperkt door het zoekformulier.

Het concept dat ten grondslag ligt aan het coderen van de data en haar structuur is de *triple*. In een triple worden data-eenheden en structuren aan elkaar verbonden. Er wordt daarom ook vaak van Linked Data (LD) gesproken. Door de data in de vorm van triples te publiceren, kunnen meer programmeurs en hun systemen gebruik maken van deze data.

1 De data op het WWW wordt daarom in het Engels het *hidden web* genomen.

Het resultaat van de export voor de Linked Data wordt “RDF” genoemd. RDF staat voor Resource Description Framework. Het drukt triples uit door aan elk data-onderdeel een URI (vaak een url) toe te kennen.

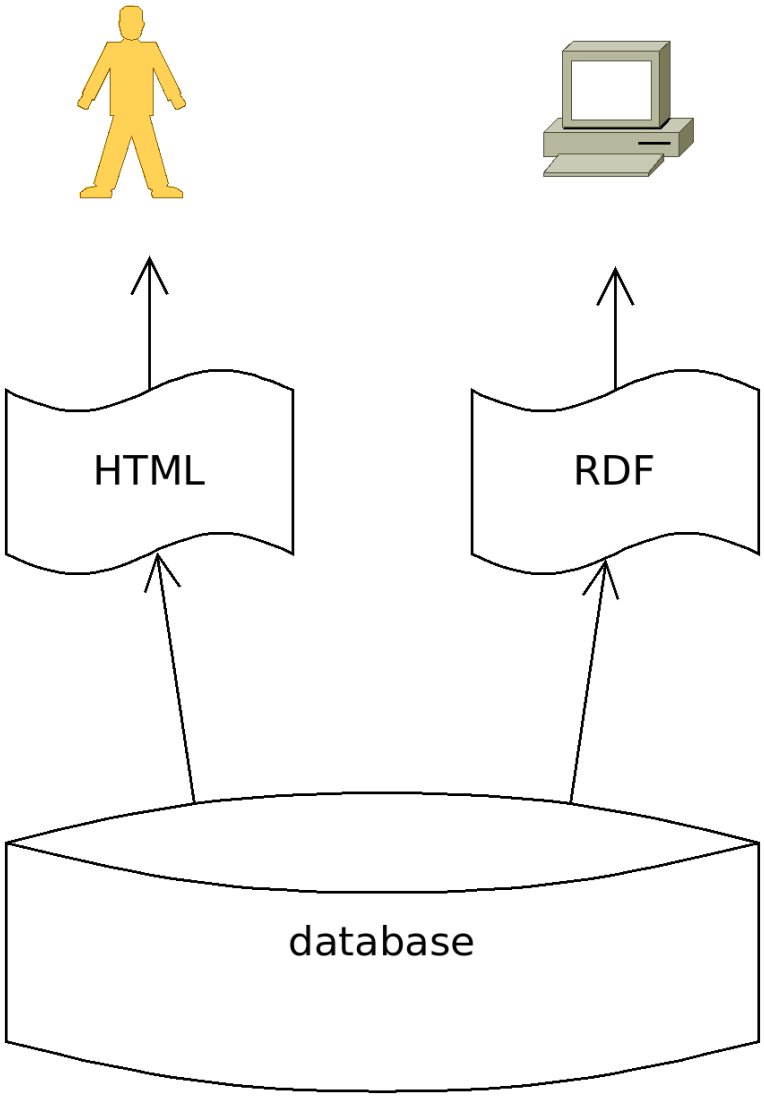
Een voorbeeld van een triple is:

“De Nachtwacht” “is geschilderd door” “Rembrandt van Rijn”.

Deze triples moeten worden gecodeerd met behulp van URLs:

<http://dbpedia.org/resource/Nachtwacht> <http://purl.org/dc/elements/1.1/creator> <http://dbpedia.org/resource/Rembrandt>.

Deze urls worden voor de leesbaarheid vaak als volgt afgekort:
<dbpedia:Nachtwacht> <dc:creator> <dbpedia:Rembrandt>



2.3. Wat levert het op?

Data publiceren op het Semantisch Web (of “Web van Data”) leidt tot meer en breder gebruik van de data, zonder extra investeringen van de erfgoedinstellingen. Men bereikt een nieuwe doelgroep, die tot nu toe onbelicht is gebleven: de programmeurs/ ICT ontwikkelaars.

Geinteresseerde programmeurs kunnen zelfstandig toepassingen ontwikkelen die te duur zijn of misschien nog niet eerder waren bedacht bij de erfgoedinstelling zelf. Als alle data voor iedereen zonder voorwaarden beschikbaar is, wordt gesproken van Linked Open Data.

Indien gewenst kan de data ook afgeschermd worden: de data wordt dan alleen verstrekt als de programmeur over de juiste inlog-codes beschikt. De programmeur kan daarna dan een eigen initiatief ontwikkelen. Omdat de data niet volledig open is, wordt niet meer van Linked *Open* Data gesproken, maar van Linked Data.

De beschikbaarstelling met behulp Semantisch Webtechnologie maakt aansluiting mogelijk bij initiatieven die nu al semantische webtechnologie gebruiken, als Europeana en zo een breder (internationaal) publiek trekken.

3. Openheid

3.1. Visie

Bij het beschikbaar stellen van data zonder beperkingen is allereerst van belang na te gaan of dit past in de strategie en visie van de organisatie.

Er zijn organisaties die bepleiten dat zij als maker van de data het recht hebben om de data onder voorwaarden beschikbaar te stellen. Deze voorwaarden kunnen dan zijn dat er voor de data moet worden betaald of dat de naam van de instelling moet worden vermeld. Ook wordt vaak het commercieel hergebruik van data niet toegestaan. Of een organisatie hiermee juridisch in haar recht staat, wordt meestal niet onderzocht.

Anderen bepleiten dat de data met behulp van belastinggeld tot stand zijn gekomen, waardoor aan hergebruik of openbaarmaking geen verdere voorwaarden kunnen worden verbonden. Volgens deze visie heeft iedereen de vrijheid om gebruik te maken van de data. Het biedt dan iedereen alle vrijheid om innovatieve projecten te doen met behulp van de data.

3.2. Rechthebbende

Voordat met de publicatie van de Linked Open Data (LOD) kan worden begonnen is het belangrijk na te gaan wie zich de eigenaar noemt van de data. Het is wenselijk met deze eigenaar in onderling overleg overeenstemming te bereiken over het gebruik van de data.

Dit rapport gaat verder niet in op de juridische aspecten van het eigendom van data. Een juridische strijd aangaan over het eigendom van data is erg ingewikkeld en levert verstoorde verhoudingen op. Desalniettemin is het belangrijk het belang van het openstellen van data te blijven afwegen tegen de eventuele juridische consequenties. Win bij twijfel advies in bij een jurist[2].

2 Meer informatie is ook beschikbaar via:
http://www.surffoundation.nl/nl/themas/digitalerechten/data/

4. Basis-stappen

Er zijn verschillende manieren om de data als Linked Open Data te publiceren. De meest basale manier wordt omschreven in dit hoofdstuk[3].

Om de implementatie eenvoudig te houden is het belangrijk een architectuur te kiezen die zo onafhankelijk mogelijk is van de omgeving die door een organisatie al is gerealiseerd. Deze bestaande omgeving wordt dan niet afhankelijk van de ontwikkeling van de publicatie van Linked Open Data. Om optimaal gebruik te kunnen maken van data, moet ook het data*model* gepubliceerd worden. Beiden worden in RDF uitgedrukt.

De onderstaande omschrijving is technisch van aard, maar geeft een indruk van wat er moet gebeuren. Het geeft een ontwikkelaar aanwijzingen om de Linked Open Data publicatie te realiseren.

4.1. Datamodel omzetten naar RDF

De meerwaarde van de publicatie van de data als LOD is dat het data*model* ook beschikbaar is. De programmeur die gebruik maakt van je data weet dan dat “Rembrandt van Rijn” de “vervaardiger” is van het schilderij, ook als hij geen kunsthistoricus is. Daar kan hij gebruik van maken. Om dit te realiseren moet het datamodel worden uitgedrukt in RDF. Hiervoor worden url's gemaakt voor elk veld in het datamodel.

Indien gebruik gemaakt kan worden van een standaard metadatamodel (zoals Dublin Core of Encoded Archival Description) dan verdient dat de voorkeur. De programmeur die de data verwerkt hoeft dan maar één keer kennis te nemen van het datamodel om ook in de toekomst bij andere vergelijkbare data-leveranciers toepassingen te kunnen maken. Sommige van deze standaard-datamodellen zijn al naar RDF vertaald. Dublin Core is daarvan een voorbeeld. EAD is (nog) niet naar RDF vertaald.

Er kan voor worden gekozen de eigen data te verschralen naar een standaard metadatamodel. Dat is echter niet nodig: in dat geval is het beter het eigen datamodel te gebruiken en in RDF aan te geven waar de overlap is met het standaard metadatamodel.

4.2. Data omzetten naar RDF

Zodra het datamodel in RDF is uitgedrukt, kan de conversie van de data zelf naar RDF worden geprogrammeerd. Bij de conversie worden url's gekozen die de onderdelen van de data identificeren. Zo zal er onder meer een url zijn voor elk object.

4.3. RDF-bestanden online zetten

De data die naar RDF is omgezet, wordt vervolgens online beschikbaar. Dit gebeurt op dezelfde manier als op het WWW met behulp van url's. Bij het volgen van elke url die is gekozen wordt een RDF-bestand getoond waarin de data is opgenomen. In de WWW-pagina waarop hetzelfde object getoond wordt, kan een verwijzing worden opgenomen

3 Lees ook: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

naar deze url, waar de data in RDF is te vinden. De verwijzing van de HTML-pagina naar de RDF-pagina, kan automatisch worden opgenomen.

4.4. Consolideren: vaste procedures op vaste servers

Voor het converteren van de data naar RDF moeten vaste afspraken worden gemaakt, zodat de data in RDF net zo actueel is als de database zelf.

Url's moeten niet meer veranderen: als iemand wil verwijzen naar jouw data, moet dat blijven kunnen.

5. Aanvullende stappen

Door de stappen te volgen zoals omschreven in hoofdstuk 4, wordt de data zó gepubliceerd dat deze door anderen kan worden gebruikt. Denk daarbij aan specialistische zoekmachines, zoals op geschiedenisvanzuidholland.nl.

Het is echter niet mogelijk direct een zoekvraag te stellen op de data die beschikbaar is gesteld. Net als op het WWW moet de data daarvoor eerst beschikbaar komen in een index of zoekmachine: een website wordt op het WWW immers pas gevonden als deze is opgenomen in de indexen van bijvoorbeeld Google.

Hetzelfde geldt voor de Linked Data. Als deze beschikbaar is in een zoeksysteem, kunnen vragen worden gesteld, zoals we dat ook gewend zijn bij databases.

5.1. Inrichten van een apart opslagsysteem voor LOD

Het stellen van een zoekvraag kan worden gerealiseerd door een opslagsysteem te maken naast het bestaande opslagsysteem. In het nieuwe opslagsysteem wordt de data in de vorm van triples opgeslagen, kan de data op verzoek worden uitgevoerd en kan een zoekvraag (“query”) worden gesteld.

Een dergelijk opslagsysteem wordt een *triplestore* genoemd.

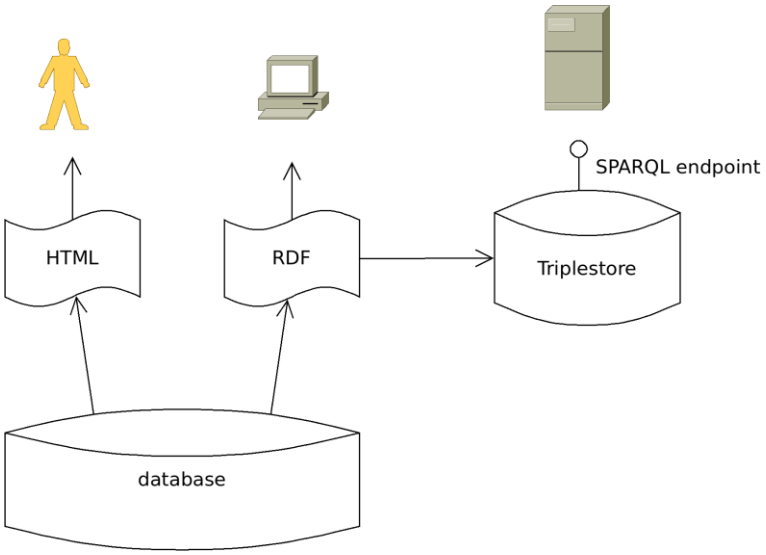
5.2. SPARQL-endpoint beschikbaar stellen

Met behulp van de triplestore bied je een zogenaamd SPARQL-endpoint aan, waaraan een programmeur zelf een query kan stellen in de speciale op RDF gerichte query taal SPARQL[4]. SPARQL is een taal voor programmeurs die goed vergelijkbaar is met SQL voor databases.

4 SPARQL (spreek uit: “Sparkel”) is de Query Language voor RDF. De afkorting staat nergens (meer) voor.

5.3. Consolideren: vaste procedures op vaste servers

De nieuwe omgeving (de triplestore) moet worden gehost en beheerd om ook na de afronding van het project deze dienst te blijven aanbieden.



6. Aanpak

De bovenstaande stappen kunnen worden gerealiseerd door drie verschillende soorten partijen:

1. iemand uit de organisatie zelf
2. een student
3. een externe leverancier

6.1. Zelf doen

Indien de organisatie beschikt over een medewerker met ICT-kennis en ervaring (die bijvoorbeeld zelf onderdelen van de website heeft geprogrammeerd), dan zou deze kunnen worden ingezet. Het publiceren van Linked Open Data zou een interessant project voor de organisatie kunnen zijn om ervaring op te doen met deze materie.

Het is belangrijk te blijven bewaken of het bouwen aan deze functies door een interne medewerker steeds de beste methode is. Wellicht komt er een moment waarop de LOD-publicatie verder moet worden geprofessionaliseerd, al dan niet door de verdere ontwikkeling alsnog uit te besteden.

6.2. Student vragen

De publicatie van LOD kan worden uitgewerkt tot een stage-project voor een technisch georiënteerde student. Deze student zal moeten worden begeleid om hem of haar kennis te laten opdoen over erfgoed.

Hierbij moet worden voorkomen dat er een oplossing ontstaat die moeilijk kan worden geconsolideerd. Bijvoorbeeld omdat er software op servers in de onderwijsinstelling werkzaam is. Het overdragen van kennis en software op de organisatie moet daarom goed worden georganiseerd.

6.3. Bedrijf uitdagen

Derde mogelijkheid is om een bedrijf uit te nodigen de publicatie van LOD te realiseren. Voor veel bedrijven zal dit iets innovatiefs zijn, waardoor er de mogelijkheid bestaat een gezamenlijk innovatief project te starten, waarbij gunstige voorwaarden kunnen worden bedongen.

Belangrijk is met elkaar tot goede afspraken te komen over kosten en inzet. Er kan tijdens de innovatie verrassend veel tijd zijn gebruikt: het is niet wenselijk daarvoor alle lasten te moeten dragen. Ook bestaat het risico dat het bedrijf zijn inzet afneemt, zodra hun interesse afneemt of andere opdrachten worden aangenomen.

6.4. Mengvorm

Natuurlijk zijn allerlei mengvormen van deze drie mogelijkheden denkbaar: een student bij een bedrijf of begeleiding van een interne medewerker door een bedrijf.