

Hoe kan de archiefsector bijdragen aan en gebruik maken van het Semantisch Web?

Handreiking Linked Data voor archieven

Versie 0.6

November 2016



Inhoudsopgave

1. Inleiding	4
2. Rationale van het Semantisch Web	5
3. Architectuur en het Semantisch Web	6
Digitaal Erfgoed Referentie Architectuur	
Model Architectuur Rijks Archiefinstellingen	
4. Linked Data terminologie	8
Vastleggen van kennis	
Principes	
Standaarden en technologie	
5. Gefaseerd met Linked Data aan de slag	13
1 - Oriënteren	
2 - Scope bepalen	
3 - Partners en leveranciers betrekken	
4 - Linked Data maken en gebruiken	
5 - Implementeer aanmaken Linked Data	
6 - Linked Data publiceren	
7 - Linked Data onderhouden	
6. Archiefsector en het Semantisch Web	19
Vierdaagse LOD pilot	
RedBot	
ECHOES 2.0	
Nationaal Archief	
7. Visie van leveranciers op het Semantisch Web	24
DEVENTit	
Spinqe	
8. Naar de toekomst	26
9. Meer kennis	27
Door te lezen	
Door te bekijken	
Door te participeren	

Leeswijzer

In **hoofdstuk 1** wordt de totstandkoming van deze handreiking binnen het innovatieprogramma Archief 2020 toegelicht.

Linked Data is een technologie, dus een middel. Het bovenliggende doel is het Semantisch Web. In **hoofdstuk 2** wordt uitgelegd wat het Semantisch Web is en wat de meerwaarde is voor de archiefsector: verbinden van collecties en aansluiten op de buitenwereld.

Archiefinstellingen die onder architectuur werken en gebruik maken van referentiearchitecturen als de DERA en MARA komen principes tegen de een verband hebben met het Semantisch Web. In **hoofdstuk 3** worden deze principes, die ingevuld kunnen worden met Linked (Open) Data, uitgelicht.

Het Semantisch Web en Linked Data komen met hun eigen terminologie. **Hoofdstuk 4** beschrijft de belangrijkste termen rondom het vastleggen van kennis en de principes, standaarden en technologie rondom Linked Data.

Archieven die zelf aan de slag willen met Linked Data kunnen gebruik maken van de in **hoofdstuk 5** beschreven aanpak zoals deze is opgesteld door Regionaal Archief Nijmegen in een Linked Data pilot.

De archiefsector zet al stappen op het vlak van het Semantisch Web. **Hoofdstuk 6** beschrijft enkele inspirerende cases waar ook nadrukkelijk wordt gekeken naar de opbrengst.

Het gebruik van Linked Data staat niet op zich. Veelal zal er een relatie zijn met bestaande beheer- en presentatie software. **Hoofdstuk 7** biedt ruimte aan leveranciers om hun visie te geven over het Semantisch Web in relatie tot hun product(en).

Hoofdstuk 8 biedt een samenvatting van de handreiking en somt enkele kritische succesfactoren factoren op.

Tot slot biedt **hoofdstuk 9** een overzicht van leesmateriaal, presentatie en communities waar u nog meer verdieping kunt vinden op het vlak van Linked Data.

1. Inleiding

De Nationale Strategie voor Cultureel Erfgoed¹ is bedoeld om de gedigitaliseerde collecties van musea, bibliotheken, wetenschappelijke instituten en archieven aan elkaar te koppelen door de ontwikkeling van een landelijke, sector overstijgende infrastructuur van voorzieningen voor digitaal erfgoed.

“Op die manier creëren slim gelegde verbanden tussen archiefgegevens heel nieuwe mogelijkheden. Voor het vertellen van nieuwe verhalen bijvoorbeeld of voor nieuw onderzoek. Zo krijgen de archieven als geheel meer waarde dan elk afzonderlijk.”

De gewenste koppelingen van gedigitaliseerde collecties moeten niet alleen gemaakt worden tussen de verschillende sectoren, maar dus ook tussen de collecties binnen de sectoren. Dat geldt zeker voor de archiefsector met daarin een grote verscheidenheid aan datasets die niet met elkaar verbonden zijn, terwijl voor Open Toegang hét uitgangspunt is: verbinden.

Het innovatieprogramma Archief 2020 heeft voor de doelstelling Toegankelijkheid van programmaliijn 3 het project “Open Toegang” geïnitieerd met daarin het onderdeel Semantiek en Linked Data. Experimenteren, kennisbevordering en kennisdelen zijn daarmee onlosmakelijk verbonden. Deze “Handreiking Linked Data” maakt daar deel van uit. Het project draagt bij aan mijlpaal 2 van het Werkplan 2016 van Archief 2020: Welke stappen overheden en hun archiefbeherende instellingen kunnen zetten om (overheids)informatie in de digitale collectie Nederland zichtbaar, toegankelijk en herbruikbaar te maken.

Op basis van bronnen over het onderwerp, reeds opgedane ervaringen, visie van leveranciers en input uit andere Archief 2020 projecten is deze handreiking samengesteld over het Semantisch Web (als doel) en Linked Data (als technologie). Het biedt de archiefsector handvatten om ermee aan de slag te gaan.

1

http://www.den.nl/art/uploads/files/Publicaties/Nationale_Strategie_Digitaal_Erfgoed_MinOCW.pdf

2. Rationale van het Semantisch Web

Het Semantisch Web is een toevoeging aan het gewone internet zoals we dat kennen – met documenten (tekst en media) die wij als mens kunnen lezen en bekijken, ook wel “Web van Documenten” genoemd. Het Semantisch Web, voegt daar betekenis aan toe. En wel betekenis aan data die gestructureerd en leesbaar is door computers. Het resultaat is een “Web van Data”.



Voor het gestructureerd vastleggen van betekenis zijn diverse methoden voorhanden, waaronder vocabulaires, ontologieën, taxonomieën en thesauri. Wanneer de data met betekenis aan elkaar wordt gerelateerd kunnen nieuwe inzichten ontstaan en nieuwe toepassingen worden gerealiseerd. Het koppelen van de eigen data, inclusief betekenis, aan andere datasets maakt de eigen data rijker (= het netwerk effect), meer bruikbaar en meer toegankelijk.

Waar het Semantisch Web gezien kan worden als doel, is Linked Data het middel waarmee dit bereikt kan worden. Een diverse set aan standaarden en technologieën, waaronder RDF en triple stores, maken het mogelijk dat Linked Data kan worden geproduceerd en geconsumeerd op een wijze waardoor verbinding tot stand komt.

Kern van het Semantisch Web is dat betekenisvolle data vrij beschikbaar is. Vaak wordt er dan ook gesproken over Linked Open Data, waarmee wordt bedoeld dat de Linked Data beschikbaar is voor hergebruik onder een open licentie. Hoewel de openheid belangrijk is voor het Semantisch Web, kan Linked Data - als middel - ook “intern” bij een organisatie gebruikt worden om data aan elkaar te linken, ook met externe bronnen, zonder dat deze “extern” wordt gepubliceerd maar waardoor wel de doorzoekbaarheid wordt verbeterd.

Veelal is er een koppeling met bestaande informatie beheer systemen binnen de organisatie die, gebruik maken van thesauri en ontologieën en dienen als basis voor het publiceren van Linked Data. Linked Data is veelal niet direct een eindproduct voor consumenten, hiervoor dient software ingezet te worden die gebruik kan maken van Linked Data. Linked Data kan gezien worden als ‘database’ die op een creatieve wijze ontsloten moet worden voor eindgebruikers. Wanneer er Linked Open Data wordt geboden kunnen ook derden toepassingen maken.

De meerwaarde van het Semantisch Web voor de archiefsector is het aan elkaar kunnen linken van data uit verschillende collecties binnen de sector én aan data die van buiten de sector afkomstig is.

Zowel data van de erfgoed sector als ook overheidsinformatie lenen zich goed om te linken met andere data, waarmee het onderdeel wordt van het Semantisch Web. Het gebruik van en bijdragen aan het Semantisch Web heeft als resultaat dat computers beter kunnen zoeken en verbanden kunnen leggen, over data sets van diverse organisaties heen. Er wordt kennis op een machine leesbare wijze verspreid.

3. Architectuur en het Semantisch Web

In diverse referentiearchitecturen worden principes benoemd die een relatie hebben met het Semantisch Web en ingevuld kunnen worden met behulp van Linked (Open) Data.

Digitaal Erfgoed Referentie Architectuur

De Digitaal Erfgoed Referentie Architectuur² (DERA) gaat uit van het strategische doel om de gebruiker centraal te stellen om op die manier het gebruik en de meerwaarde van het digitale erfgoed in Nederland te vergroten. Om deze doelen in Netwerk Digitaal Erfgoed (NDE) verband te kunnen realiseren zijn een zestal principes opgesteld die aangeven hoe de erfgoedsector hiervoor zorg dient te dragen:

- authenticiteit van erfgoed informatie is duidelijk
- beschikbaarheid van erfgoed informatie is duidelijk
- erfgoed informatie is herkenbaar en wordt gebruiksvriendelijk aangeboden
- erfgoed informatie is eenduidig beschreven
- erfgoed informatie verwijst
- erfgoed informatie is verwijfsbaar

Met name de laatste 3 principes laten zich vertalen naar de toepassing van Linked Data. Dit biedt diverse uitwisselingsformaten en notaties waarmee (de betekenis van) erfgoed informatie beschreven kan worden. Door vanuit de erfgoed informatie links te leggen met andere Linked Data bronnen zoals Geonames en gebruik te maken van thesauri komt er meer betekenis en wordt er verwezen naar andere informatie. Wanneer de erfgoed informatie ook als Linked Data gepubliceerd wordt, kunnen anderen hiernaar verwijzen.

Het NDE voorziet een geleidelijke overgang waarbij erfgoedinstellingen zoveel mogelijk gefaciliteerd en gestimuleerd zullen worden om onderdeel te worden van het groeiende netwerk van gedistribueerde erfgoed informatie. Op nationaal niveau zal het NDE investeren in landelijke voorzieningen die erfgoedinstellingen helpen bij het publiceren van hun datasets als Linked Data, bijvoorbeeld via het realiseren van het register en landelijke terminologiebronnen voor wie, wat, waar en wanneer en de knowledge graph voor de integratie van erfgoed informatie.

2

http://www.den.nl/art/uploads/files/Publicaties/20160331_NDE_DEF_PositionPaper_NetwerkErfgoedInformatie.pdf

Model Architectuur Rijks Archiefinstellingen

De Model Architectuur Rijks Archiefinstellingen³ (MARA) is een gedeelde visie van Convent-RHC's en Nationaal Archief en beschrijft wat er nodig is aan organisatie, applicaties en informatie, infrastructuur en beveiliging om deze visie te realiseren.

In de MARA worden ook principes voor dienstverlening en voor de uitwisselbaarheid en verbindbaarheid van data en meta data opgesomd. De principes waar het Semantisch Web en Linked Data een rol kunnen spelen zijn hieronder integraal overgenomen:

Naam	Documentatie
MARA Semantische en technische verbinding mogelijk	Het is mogelijk om elementen uit collecties semantisch aan elkaar te relateren en ook technisch aan elkaar te verbinden zodat ze in samenhang gepresenteerd kunnen worden.
MARA Zoeken en vinden via verbindingslaag	Zoekinterfaces op een publiekswebsite hebben geen directe interactie met een collectie beherend systeem. Elke zoekopdracht verloopt via een aggregator, waardoor het altijd mogelijk is om meta data uit meerdere collectie beherende applicaties aan elkaar te verbinden. Dit kunnen bijvoorbeeld meerdere collecties binnen één instelling zijn, of één type collectie over meerdere instellingen (zoals bij de Rijkscollectie).
MARA Data en meta data beschikbaar als open data	Waar mogelijk - behoudens (wettelijke) beperkingen - stellen collectie beherende instellingen meta data en data beschikbaar als open data, via een open toegang en op een gestandaardiseerde manier.
MARA Gebruik van kernregistraties	<p>Bij het beschrijven (meta dateren) van (archief-)materiaal wordt maximaal gebruik gemaakt van al bestaande en liefst gedeelde kernregistraties / authority files. Oftewel: liever verwijzen dan beschrijven.</p> <p>We onderscheiden twee hoofdentiteiten waarvoor aparte registraties bestaan: Archief (= NEN ISO 23081 Record) en Zorgdragers / Archiefvormers (= NEN ISO 23081 Actor)</p> <p>N.B. We onderscheiden geen "basisregistraties" voor de andere in NEN ISO 23081 genoemde entiteiten, zoals Mandaat en Activiteit.</p>
MARA Gebruik domein specifieke beschrijvingsstandaarden	Voor entiteiten voor het beschrijven waarvan al standaarden aanwezig zijn, gebruiken we zoveel mogelijk deze standaarden. Dit geldt zowel voor het beschrijven van objecten als voor het exporteren van de beschrijvingen

³ http://www.nationaalarchief.nl/sites/default/files/docs/mara_1_0.pdf

4. Linked Data terminologie

Linked Data is een manier van publiceren van data op het Semantisch Web zodat:

- hergebruik wordt gestimuleerd
- redundantie afneemt
- verbondenheid wordt gemaximaliseerd
- er waarde toevoeging aan data ontstaat door netwerk effect

In dit hoofdstuk worden de Linked Data principes en bijbehorende terminologie uiteen gezet.

Vastleggen van kennis

Een (gecontroleerde) **vocabulaire**, of **termenlijst**, is een lijst van termen die expliciet zijn opgesomd. Deze lijst wordt gecontroleerd en beschikbaar gesteld door een autoriteit. Alle termen in een gecontroleerde vocabulaire dienen een eenduidige, niet-redundante definitie te hebben.

Een (formele) **ontologie** is een gecontroleerde vocabulaire uitgedrukt in een ontologie taal. Deze taal heeft een grammatica voor het gebruik van de woordenschat om iets zinvols binnen een bepaald domein van belang uit te drukken. De grammatica bevat formele beperkingen over de manier waarop termen in gecontroleerde vocabulaire van de ontologie van elkaar kunnen worden gebruikt.

Voorbeelden:

- DC⁴ (Dublin Core) is een sector overstijgende vocabulaire waarmee online bronnen van meta data kunnen worden voorzien.
- FOAF⁵ (Friend of a Friend) is een machine-leesbare ontologie, die personen beschrijft, de relaties met andere personen en voorwerpen en de interacties hiervan.
- SKOS⁶ (Simple Knowledge Organization System) is een standaard van het World Wide Web Consortium (W3C) waarmee zowel de structuur als de onderlinge relaties voor woordenlijsten, thesauri en taxonomieën kan worden beschreven en vastgelegd.
- BAG⁷ (Basisregistratie Adressen en Gebouwen) vocabularium.
- Geonames⁸ is een geografische database.

⁴ <http://dublincore.org/>

⁵ <http://www.foaf-project.org/>

⁶ <http://www.w3.org/TR/skos-primer/>

⁷ <http://lod.geodan.nl/vocab/bag/>

⁸ <http://www.geonames.org/>

Taxonomie, of **classificatieschema**, is de wetenschap van classificatie. Het verwijst naar de indeling van de dingen of concepten en de schema's die ten grondslag liggen aan een dergelijke kwalificatie. Een taxonomie heeft normaal gesproken enkele hiërarchische relaties ingebed in de classificaties.

Een **thesaurus** kan worden opgevat als een uitbreiding op een taxonomie: onderwerpen worden gerangschikt in een hiërarchie en bovendien is er de mogelijkheid om andere uitspraken worden gedaan over de onderwerpen zoals "bredere term" en "nauwere term".

Voorbeelden:

- Erfgoedthesaurus⁹, een centrale laag met cultuurtermen, die alle collectiegegevens van deelnemende erfgoedinstellingen met elkaar verbindt tot een breed kennisnetwerk beheerd door de Rijksdienst.
- GTAA¹⁰ (Gemeenschappelijke Thesaurus Audiovisuele Archieven), de thesaurus van Beeld en Geluid
- AAT-NED¹¹ (Art and Architecture Thesaurus - Nederlandstalig), is een thesaurus voor cultuur- en erfgoedterminologie van het Amerikaanse Getty Research Institute.
- NTA¹² (Nederlandse Thesaurus van Auteursnamen), ook wel de Persoonsnamenthesaurus genoemd, is onderdeel van het Gemeenschappelijk Geautomatiseerd Catalogiseersysteem.

Overzichten van termenlijsten en thesauri:

- Overzicht terminologiebronnen website DEN, Kenniscentrum digitaal erfgoed¹³
- Overzicht thesauri en terminologiebronnen website Art & Architecture Thesaurus – Nederlandstalig¹⁴
- Identification of existing terminology resources in museums / Marie-Véronique Leroi (MCC/MRT), Johann Holland (Michael Culture Aisbl). – Athena, 2009¹⁵
- BARTOC¹⁶ (Basel Register of Thesauri, Ontologies & Classifications).

⁹ <http://www.erfgoedthesaurus.nl>

¹⁰ <http://gtaa.beeldengeluid.nl/>

¹¹ <http://browser.aat-ned.nl/>

¹² <https://www.oclc.org/support/services/ggc/nta.en.html>

¹³ <http://www.den.nl/terminologiebronnen>

¹⁴ <http://website.aat-ned.nl/Andere%20thesauri>

¹⁵

http://www.packed.be/images/uploads/projects/D4.1_Identification_of_existing_terminology_resources_in_museums.pdf

¹⁶ <http://www.bartoc.org/>

Principes

Tim Berners-Lee beschreef¹⁷ in 2007 de 4 principes van Linked Data:

1. Gebruik **URI's** om dingen te identificeren.
2. Gebruik **HTTP URI's** zodat er naar deze dingen kan worden verwezen en dat ze kunnen worden opgezocht.
3. Leg de informatie over het ding vast in een triple in **RDF**.
4. Neem **links** op naar andere URI's zodat gerelateerde dingen gevonden kunnen worden.

Hierin staat:

- **URI** voor uniform resource identifier, dit is een eenmalige benaming van een "bron", een stuk informatie, data of dergelijke. Een URL is een voorbeeld van een URI.
- **HTTP** voor Hypertext Transfer Protocol, dit is het protocol voor de communicatie tussen een web client en een webserver.
- **RDF** voor Resource Description Framework, dit is een formaat om gegevens in het algemeen voor te stellen en uit te wisselen. In de paragraaf "Standaarden en technologie" wordt hier nadere op in gegaan.

Om partijen die informatie publiceren te stimuleren om hun data in een zo herbruikbaar mogelijk formaat beschikbaar te stellen, heeft Tim Berners-Lee een vijfsterrenmodel voorgesteld. Hierbij worden de volgende sterren toegekend:

-
- ★ 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 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.

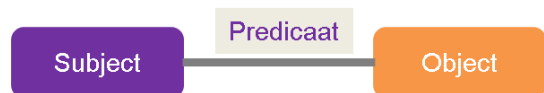
5 sterren Linked Open Data vormen het essentiële onderdeel van het Semantisch Web!

¹⁷ <http://www.w3.org/DesignIssues/LinkedData.html>

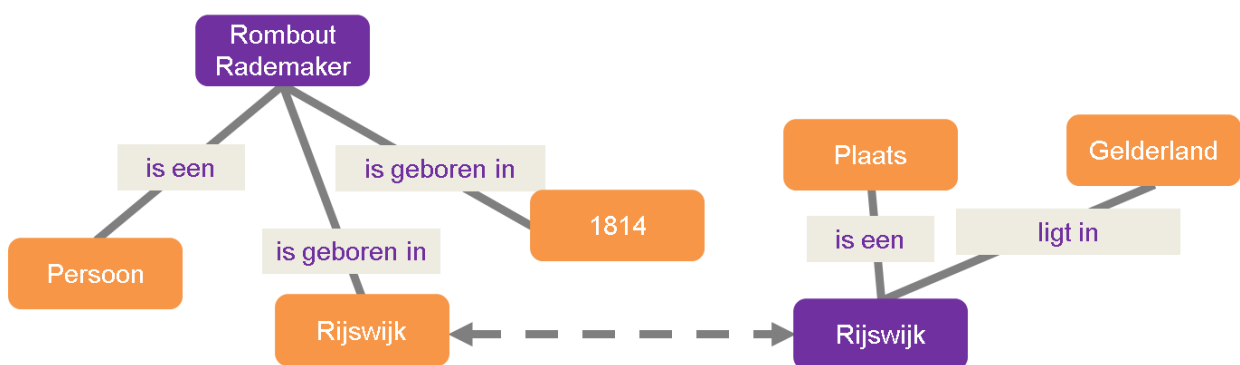
Standaarden en technologie

- **Resource Description Framework**¹⁸ (RDF) is een standaard van het World Wide Web Consortium (W3C), oorspronkelijk ontworpen als een meta data-model, maar gaandeweg gebruikt als een formaat om gegevens in het algemeen voor te stellen en uit te wisselen. Het onderliggende RDF datamodel kent meerdere syntaxes (of serialisatie formaten), bijvoorbeeld RDF/XML, N3, Turtle, N-Triples, Trig, Trix.

Met het RDF-model kunnen uitspraken gedaan worden over de kenmerken van bronnen op het web (resources) in de vorm van een driedelige subject-predicaat-object-structuur, of in RDF-terminen een **triple**. Het subject is in essentie de resource die beschreven wordt. Het predicaat is welk kenmerk of aspect van die bron beschreven wordt. Het object ten slotte is wat de waarde van dat kenmerk is.



In onderstaand schema worden 6 triples weergegeven, waarbij er alleen tekst wordt getoond. In de praktijk zullen subject en predicaat bestaan uit URI's, het object kan een URI of tekst zijn.



In het voorbeeld kan niet gezegd worden dat de twee Rijswijken hetzelfde zijn (aangeduid met de stippellijn), er is onvoldoende kennis om vast te stellen of beide Rijswijken dezelfde “dingen” zijn en of het de Rijswijk in Gelderland of Zuid-Holland of Noord-Brabant betreft. Wanneer er in de triples een URI was opgenomen voor de plaats, bijv. <http://sws.geonames.org/2748076/> (de URI beschrijft Rijswijk in Zuid-Holland bij Geonames ontologie), dan was de koppeling wel eenduidig (en geautomatiseerd) te leggen.

Dit voorbeeld toont aan dat het gebruik van vocabulaires, ontologieën, taxonomieën, thesauri zeer belangrijk is voor het koppelen van data. Wanneer de kennis over de eigen data eenduidig wordt vastgelegd wordt het linken en beter ontsluiten mogelijk gemaakt.

¹⁸ https://nl.wikipedia.org/wiki/Resource_Description_Framework

- De **Web Ontology Language**¹⁹ (**OWL**) is een W3C-standaard voor het definiëren van web ontologieën. Deze ontologieën worden gekenmerkt door de formele semantiek en RDF / XML-gebaseerde serialisaties voor het Semantisch Web.
- **RDF Schema**²⁰ (andere afkortingen: RDFS, RDF(S), RDF-S, of RDF/S) is een set van klassen met bepaalde eigenschappen van de representatietaal RDF. Het voorziet in basiselementen voor de ontologie, ook wel "RDF-vocabulaire" genoemd, om structuur te geven aan RDF-bronnen. Deze RDF-bronnen kunnen met de zoektaal SPARQL worden geraadpleegd.
- **SPARQL** (Simple Protocol And RDF Query Language) is een RDF zoektaal die gebruikt wordt om RDF-gebaseerde data te bevragen middels zoekopdrachten/queries.
- Een **triple store** of RDF store is een special database voor de opslag en opvragen van triples.

¹⁹ https://en.wikipedia.org/wiki/Web_Ontology_Language

²⁰ <https://nl.wikipedia.org/wiki/RDFS>

5. Gefaseerd met Linked Data aan de slag

De basis van dit hoofdstuk is gelegd door het document “Stappenplan Linked (Open) Data voor Archieven”, gemaakt²¹ door het Regionaal Archief Nijmegen als resultaat van hun Linked Data pilot, aangevuld met feedback uit de archiefsector. In het volgende hoofdstuk worden de resultaten van deze Vierdaagse LOD pilot beschreven.

1 - Oriënteren

“Wat is Linked Data?”

Verzeker je ervan dat je goed op de hoogte bent van wat Linked Data is en wat het brengt aan extra toepassingsmogelijkheden voor de collectie. Dit kan enerzijds door het lezen van publicaties of bijwonen van presentaties. Anderzijds helpt het om gerichte cursussen op dit terrein te volgen om de nuances in de belofte van Linked Data te ontdekken. Kritische vragen kunnen worden beantwoord door iemand uit de praktijk, je medecursisten stellen bovendien vaak andere vragen zodat je elkaars kennisbasis vergroot. Naast het nut van de toepassing van Linked Data moet de Linked Data terminologie bij alle projectleden bekend zijn. Zoek ook zeker aansluiting bij relevante communities en zoek collega instellingen op om kennis te delen.

2 - Scope bepalen

“Voor wie is de Linked Data?”

Als je overtuigd bent van de meerwaarde van Linked Data voor je collectie stel dan vast wie de doelgroep is, wat de use cases zijn en de business case is. De doelgroep kunnen externe partijen zijn, maar de Linked Data technologie kan ook worden toegepast binnen de organisatie of binnen een beperkte groep organisaties. Hoe gaat de eigen organisatie de nieuwe kennisbron gebruiken? Probeer specifiek te zijn en te anticiperen hoe deze doelgroep de Linked Data zou kunnen gebruiken. Probeer ook in te schatten wat deze doelgroep nodig heeft om een toepassing te kunnen maken met jouw Linked Data.

Als je gegevens zich lenen voor publicatie als Linked Open Data volg dan eerst het stappenplan Open Data²².

Bepaal of je je gegevens alleen wilt meta dateren met een Linked Data thesaurus, of dat je alle gegevens in het Linked Data formaat wilt omzetten. Het gebruik van termenlijsten is (zelfs los van Linked Data) een goede eerste stap om de kwaliteit van de eigen data te verhogen. Daarnaast wordt de weg opengezet om data te linken. Bedenk hierbij dat dit linken veelal kennis vereist, kennis die bij de eigen instelling aanwezig is en minder/niet bij externe partijen als aggregatoren.

²¹ <https://informatie2020.pleio.nl/discussion/view/45253692/stappenplan-linked-data-voor-archieven-open-voor-feedback>

²² http://archiefwiki.org/wiki/Stappenplan_Open_Data

Een algemene regel is: hoe breder de doelgroep, hoe hoger de totale publicatiekosten²³. Dit heeft te maken met de extra aanpassingen die je moet maken om de Linked Data dienst beschikbaar te stellen aan gebruikers die niet bekend zijn met deze techniek. Onderstaande afbeelding plaatst de verschillende technieken op de assen publicatiekosten en gebruik.



Voor de technieken die in het linker onder kwadrant worden genoemd is specialistische Linked Data kennis nodig. Niet alle organisaties willen hierin investeren. De technieken in het rechter boven kwadrant zijn vrij algemeen bekend bij ontwikkelaars. Om deze algemene diensten te kunnen aanbieden moeten zowel de diensten van het linker onder kwadrant als de diensten van het rechter boven kwadrant worden gemaakt (de diensten links onder zijn een voorwaarde voor de diensten rechts boven).

Er geldt verder: hoe beter de datakwaliteit, hoe hoger de publicatiekosten. Voor het maken van Linked Data van een willekeurige bron zijn een aantal stappen nodig, die meer of minder tijd kosten, afhankelijk van de staat van de bron en het gewenste eindresultaat. De keuzes die je daarvoor vooraf moet maken zijn weergegeven in de volgende afbeelding.

²³ Met publicatiekosten worden alle kosten bedoeld die gemaakt worden bij het proces van het transformeren van bestaande data naar gepubliceerde Linked Data.



De onderwerpen in bovenstaande afbeelding worden hieronder apart behandeld.

- **Schone/vuile bron data:** net zoals bij andere conversies geldt bij Linked Data ook: “garbage in is garbage out” en als je dit bij het maken van Linked Data wilt herstellen kost dat (veel) tijd. Zorg dat je bron schoon is en blijft door dit bij de invoer van de gegevens in een archiefbeheersysteem goed te regelen (bijv. door gebruik termenlijsten).
- **Vocabulaires:** een vocabulaire is een datamodel dat wordt gebruikt om de Linked Data gegevens te beschrijven. Er is op het internet al een groot aantal vocabulaires beschikbaar. Hergebruik van deze vocabulaires heeft als voordeel dat je meteen interoperabel bent met bronnen die ook met dit vocabulaire zijn vastgelegd. Gebruik van meerdere en overlappende vocabulaires in je dataset zorgt ervoor dat je dataset door meerdere doelgroepen kan worden gebruikt. Je kunt ook besluiten om zelf een eigen vocabulaire in RDFS/OWL te maken. Voor een uitgebreid dataset (met veel tabellen en eigenschappen) kost dit veel tijd en je bent bovendien niet interoperabel met anderen. Doe dit alleen als (delen van) jouw domein nog niet beschreven is/zijn in een bestaand datamodel. Zie hiervoor de online catalogus van vocabulaires Linked Open Vocabularies²⁴.

Veel erfgoedinstellingen hebben hun collecties al via een thesaurus-in-Linked-Data-formaat (=SKOS) gemetadateerd. Het resultaat hiervan is dat de meta data van de collectie via het Linked Data principe aan elkaar gekoppeld is. Zo kun je dus via een trefwoord uit de Linked Data catalogus items uit collecties van meerdere instellingen vinden. Een voorbeeld van zo'n Linked Data thesaurus is de GTAA²⁵ van het Nederlands Instituut voor Beeld en Geluid en de erfgoedthesaurus²⁶ van de Rijksdienst voor het Cultureel Erfgoed.

²⁴ <http://lov.okfn.org/dataset/lov/>

²⁵ <http://gtaa.beeldengeluid.nl/>

²⁶ <http://www.erfgoedthesaurus.nl/>

- **Koppelingen:** voor het maken van Linked Data is het niet noodzakelijk dat je een koppeling met een andere bron maakt. Als je geen koppelingen maakt, heb je zogenaamde 4 sterren Linked Data. Koppelingen maken je dataset wel interessanter, rijker. Je voegt waarde toe aan je dataset. Het maken van een koppeling is een tijdrovende zaak. In de eerste plaats omdat een domeinexpert de koppeling moet valideren. Ook als de koppeling is “ontdekt” door een automatisch proces! Niet alle bronnen worden op een manier aangeboden die het makkelijk maakt om automatisch een relatie te ontdekken (bijvoorbeeld via een gelijk label). Betekenisvolle koppelingen zijn meer waard dan betekenisarme koppelingen. Een voorbeeld van een betekenisvolle koppeling is: bron <heeft als publiceerprincipe> doel. Een betekenisarme koppeling is bijvoorbeeld: bron <zie ook> doel. (De betekenis van de koppeling is tussen vishaken aangegeven). Er is voor een betekenisvolle relatie (bijvoorbeeld “heeft publiceerprincipe”) veel meer voorafgaand onderzoek nodig, dan een betekenisarme verwijzing (“zie ook”) die slechts een vage richting aangeeft van een relatie.

Kortom: het zoeken naar bronnen om mee te verbinden en het verbinden zelf is een tijdrovende puzzel, maar de moeite waard.

3 - Partners en leveranciers betrekken

“Linked Data maak je samen!”

De meest waardevolle koppeling (lees: je kunt er het beste mee automatiseren) kan alleen met andere Linked Data gemaakt worden, zogenaamde RDF-RDF koppelingen. Een koppeling tussen een Linked Data bron en een HTML pagina is minder bruikbaar. Dus het is in ieders belang dat er meerdere Linked Data bronnen komen.

Verder: stem je publicatiemethode af met je doelgroep voor de beste kosten/baten afweging. Zie hiervoor nogmaals de afbeelding over de publicatiekosten versus het gebruiksgemak in de beschrijving van de vorige fase. Maak het niet te mooi, want je organisatie draait voor de kosten op, maar ook niet te ingewikkeld, want dan wordt je dataset niet gebruikt en is alles voor niets geweest.

Als je hebt bepaald wie je doelgroep is, dan kun je daar waarschijnlijk vast contact mee maken. Kijk ook wat de andere partij voor jou kan betekenen. Hebben ze een dataset die voor jou interessant is om aan te koppelen?

4 - Linked Data maken en gebruiken

“Doen is leren!”

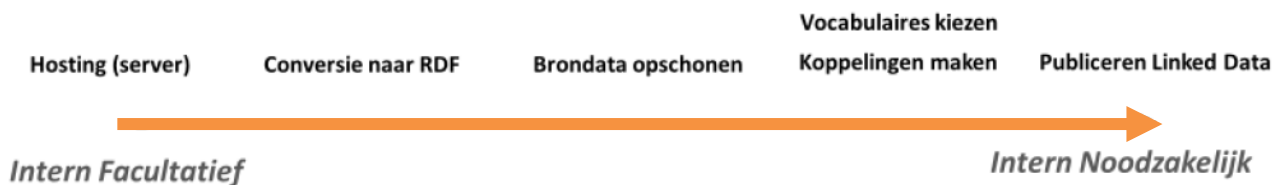
Bij het samenstellen van het Linked Data projectteam is het raadzaam verschillende deskundigen op te nemen. In het team zitten uiteraard mensen met Linked Data kennis en kennis van web technologie, die waarschijnlijk ook handig zijn met tools. Maar het is ook erg belangrijk dat er mensen in het team zijn opgenomen die inhoudelijke kennis hebben van het onderwerp van de dataset. Dit is belangrijk voor het kiezen van het juiste vocabulaire en het mappen van de bron naar de Linked Data bestemming.

Het maken van Linked Data kun je in twee stappen oppakken:

- Maak RDF van je data: kies vocabulaires en zet de data om naar triples
- Link de data met andere bronnen (handmatig of geautomatiseerd): link de data aan thesauri, GeoNames, DBpedia, enzovoort.

Begin met een eenvoudige dataset (=weinig tabellen en attributen) die relatief veel waarde heeft (=het is interessant om er aan- of mee te koppelen). Gebruik voor Linked Data bijvoorbeeld het Europeana Datamodel²⁷ en de Dublin Core²⁸ vocabulaires. Gebruik voor het meta dateren de thesauri van Beeld en Geluid of het RCE.

Niet alle dingen hoeft je zelf te doen, maar sommige stappen moet je liever niet uitbesteden. Zie onderstaande “Zelf doen of uitbesteden” afbeelding:



Het is voor je organisatie van weinig toegevoegde waarde als je de server, waarmee de Linked Data serveert, zelf host. Het is wel noodzakelijk dat je de Linked Data vanaf je eigen web domein serveert, omdat zo duidelijk wordt dat je de bronhouder bent! (Dus webadres van je RDF Resource begint met <http://data.mijnarchief.nl/> waarbij “mijnarchief” een willekeurig archief is. Het letterlijke gebruik van het sub domein “data” is niet noodzakelijk).

Als de bronbeschrijving en de mapping tussen bron en bestemming bekend is kan de conversie ook door een externe partij worden uitgevoerd. Net als het opschonen van de bron, zolang een eigen, inhoudelijk expert maar heeft aangegeven hoe de bron moet worden opgeschoond.

In het uitgebreide stappenplannen van het Platform Linked Data Nederland²⁹ en DEN³⁰ staat uitgelegd hoe je Linked Data maakt.

²⁷ <https://raw.githubusercontent.com/europeana/corelib/master/corelib-edm-definitions/src/main/resources/eu/rdf/edm.owl>

²⁸ <http://dublincore.org/schemas/rdfs/>

²⁹ <http://www.pilod.nl/wiki/BoekTNO/stappenplan>

³⁰ <http://www.den.nl/art/uploads/files/Tutorial%20Linked%20Data%20-%20Stap%201%20tm%206.pdf>

5 - Implementeer aanmaken Linked Data

“We doen het!”

Denk na over processen en systemen binnen je organisatie. Daarbij hoort de keus: wie gaat wat doen en in welk systemen ga je dat doen: collectiebeheersystemen, beeldbank of een apart Linked Data systeem. Je zult bijvoorbeeld het gecontroleerd handmatig linken waarschijnlijk door je collectiebeheerders vanuit je beheersysteem willen doen, maar automatisch linken misschien liever vanuit een triple store. Dit alles dient onderdeel te worden van het normale werkproces en dient te passen binnen je informatiearchitectuur.

6 - Linked Data publiceren

“We zijn live!”

Linked Data kan intern beschikbaar zijn of via besloten netwerken voor partners, of het kan open aangeboden worden als Linked Open Data. In alle gevallen geldt: maak bekend dat je Linked Data hebt!

Met de publicatie biedt je andere partijen de mogelijkheid om links te leggen naar jouw Linked Data en om de Linked Data te gebruiken in nieuwe toepassingen. Het promoten van de gepubliceerde kennisbron stimuleert dat de meerwaarde van Linked Data tot zijn recht komt.

Je kunt om de vindbaarheid te vergroten je dataset publiceren in een Open Data catalogus zoals data.overheid.nl of de data hub³¹. Hierin kun je aangeven dat je dataset het Linked Data formaat heeft.

7 - Linked Data onderhouden

“En verder...”

Als de gepubliceerde Linked Data door de tijd heen veranderd moet worden omdat het niet meer actueel is, of uitgebreid moet worden, dan zul je de gegevens via een redactie-publicatieproces moeten bijhouden. Hiervoor bestaan Linked Data content-management systemen. Vergeet niet een koppeling te maken met de Linked Data van eerdere versies. Het bijhouden van alle versies van je Linked Data en de verbindingen tussen de versies kan ingewikkeld en tijdrovend, zorg dus dat dit onderdeel uitmaakt van het geautomatiseerde werkproces.

³¹ <https://datahub.io/>

6. Archiefsector en het Semantisch Web

In de archiefsector zijn er diverse projecten waar het Semantisch Web, dan wel Linked Data, een belangrijke rol spelen. Enkele van deze projecten zijn hier ter inspiratie opgetekend.

Vierdaagse LOD pilot

Het Regionaal Archief Nijmegen (RAN) heeft de 100e Nijmeegse Vierdaagse aangeprezen om door middel van een pilotproject historische gegevens van dit grootste wandelevenement ter wereld als Linked Open Data te publiceren.

Het RAN beheert diverse archieven en collecties over de Vierdaagse. De gegevens zijn door middel van Semantic Wiki voor een deel gesemantiseerd en kunnen als Open Data worden hergebruikt. In al deze gevallen ging het om losse datasets, die inhoudelijk veel relaties hebben maar niet aan elkaar gekoppeld zijn. De techniek van Linked Open Data maakte het mogelijk informatie uit en over verschillende collecties, ook van archiefinstellingen onderling, met elkaar te verbinden, waardoor data kunnen worden verbeterd en verrijkt. In de pilot van het RAN is gebruik gemaakt van algemene datasets als Geonames en DBpedia.

Opbrengst

Het uiteindelijke, externe resultaat is een triple store³² met gekoppelde Vierdaagse data waarmee ontwikkelaars aan de slag kunnen. Intern is er door de pilot veel geleerd over Linked Data, welke ook tot uiting is gekomen in een stappenplan welke onderdeel vormt van deze handreiking.

Met de kennis van deze nieuwe technologie kunnen Linked Data oplossingen/initiatieven beter beoordeeld worden en kan er inhoudelijk over meegepraat worden. De pilot leverde het inzicht op dat Linked Data geen eindproduct is voor het brede publiek, maar gezien moet worden als een halffabricaat waar toepassingen gebruik van kunnen maken.

De investeringen in Linked Data liggen volgens het RAN op dit moment hoger dan de opbrengsten, wat dan ook de vraag oproept wat de rol van archiefinstellingen in de verrijking is. Gaat elke instelling een eigen triple store (techniek) opzetten en al haar data zelf verrijken (organisatie)? Of is het handiger dit op het niveau van een aggregator te positioneren? Als voorbeeld hiervan kan RedBot genoemd worden (deze wordt in de volgende paragraaf beschreven) en of de route via Archives Portal Europe (APE) via welke data van instellingen naar via Europeana als Linked Open Data in het Europeana Data Model beschikbaar kan worden gesteld. Ook regionale initiatieven als Brabant Cloud kunnen hierin een rol spelen. De Brabant Cloud³³ is een duurzame digitale infrastructuur, die op een semantische manier zorgt voor het vastleggen, bewaren en delen van alle culturele erfgoed informatie van de provincie Noord-Brabant.

³² <http://www.vierdaagselod.nl/>

³³ <http://www.brabantcloud.nl/>

Verrijking van de eigen dataset door het gebruik van externe thesauri zoals Geonames voor plaatsen in het eigen beheersysteem, wat vaak als extra module beschikbaar is, kost veel handwerk. Hulpmiddelen als de ErfGeo-verrijkingsservice³⁴ van de Digitale Collectie kunnen hierbij helpen, maar het toevoegen van kennis blijft ook mensenwerk. Niet alleen voor bestaande data zou er een conversie / toevoeging moeten plaatsvinden, ook digital born materiaal dat het e-depot binnenkomt, bevatten geen URI's dus zullen ook verrijkt moeten worden!

RedBot

Fries erfgoed hub

RedBot³⁵ wil dé online plek waar informatie is te vinden over digitaal Fries erfgoed, waaronder oude kaarten, atlanten, schilderijen, museale collecties, films en foto's, monumenteninformatie, juridische stukken, notariële akten, Friese literatuur en publicaties in tijdschriften en kranten, prenten en tekeningen. De doelgroep van RedBot bestaat uit collectie beherende instellingen en afnemers van data.

Vanuit het perspectief van RedBot bestaat de noodzaak tot integratie van informatie uit verschillende datacollecties van Friese culturele erfgoedinstellingen. Immers, het Friese digitale cultureel erfgoed is verspreid over diverse datacollecties. Het bij elkaar brengen van deze datacollecties levert meer dan alleen een compleet en gedetailleerd beeld op. Het levert ook informatieve meerwaarde door wederzijdse aanvulling van informatie uit verschillende databronnen (synergie).

RedBot verbindt de "Collectie Friesland" via Linked Data

Het integreren van data uit verschillende bronnen is een ontwikkeling die een vlucht heeft genomen door de mogelijkheden die internet biedt. De ontwikkeling van RDF en Linked Open Data is het antwoord op deze problematiek. Door middel van het 'losser' structureren van databases en het gebruik van internetadressen als 'identifiers', zogenaamde URI's, ontstaan mogelijkheden voor het flexibel verbinden ofwel linken van data uit verschillende bronnen.

Als objecten in datacollecties beschikbaar worden gesteld aan derden is het belangrijk dat de betekenis (semantiek) van de data duidelijk is omschreven in een datamodel. In RDF wordt het datamodel beschreven als ontologie waarbij zo veel mogelijk gebruik wordt gemaakt van gestandaardiseerde ontologieën en vocabulaires, zodat semantiek van de data kenbaar en dus bruikbaar is. Het is niet noodzakelijk en zelfs inefficiënt om een object zoals een plaats opnieuw te definiëren als daar al een bruikbare definitie van beschikbaar is in een breed gedeelde ontologie, zoals bijvoorbeeld schema.org dat biedt.

³⁴ <http://erfgeo.data.digitalecollectie.nl/>

³⁵ <http://www.redbot.frl/>

De beschrijving van het Fries cultureel erfgoeddomein is deels al gedaan in Wikipedia en beschikbaar als semantische database via DBpedia. Deze bronnen bieden een rijkdom aan meta data voor de beschrijving van digitaal Fries cultureel erfgoed. Het koppelen van deze bronnen levert contextuele informatie aan culturele erfgoedobjecten. DBpedia is de semantische database die is gebaseerd op Wikipedia, wat betekent dat de verschillende feiten over een onderwerp die op één Wikipedia-pagina worden genoemd als afzonderlijke feiten worden opgeslagen en toegankelijk zijn gemaakt via DBpedia.

Leveranciers van data aan RedBot, of beter gezegd deelnemers in het netwerk, beschikken meestal al over een (relationele) database waarin datacollecties worden beheerd. Om aansluiting te vinden bij de hub is het noodzakelijk die data in RDF beschikbaar te stellen. Dat kan decentraal door middel van de toevoeging van SPARQL service die toegang biedt tot een relationele database. Een leverancier kan op deze wijze eenvoudig zijn systeem 'aansluiten' op het netwerk van RedBot. In de SPARQL service die aansluit op het database systeem wordt een link gedefinieerd tussen het interne datamodel en de ontologie van RedBot. Op deze wijze wordt de interne semantiek van een datasysteem uitgedrukt in de semantiek van RedBot. Een duidelijke definitie van de ontologie van RedBot maakt het mogelijk deze afstemming decentraal uit te voeren.

Opbrengst

Het aanbod van Linked Open Data in RedBot is gebaseerd op de datacollecties van aangesloten databronnen van leveranciers, op de modellering van het Friese erfgoeddomein in de vorm van ontologie en op Linked Open Data op het Internet in bronnen als Wikipedia en DBpedia. Door de combinatie van data ontstaat een fijnmazig netwerk van informatie over het Friese erfgoed; Informatie afkomstig van Friese erfgoedinstellingen aangevuld met informatie afkomstig van andere bronnen op internet en het publiek (user generated content) eveneens.

Het aanbod van Linked Open Data aan consumenten en zakelijke afnemers biedt de mogelijkheid om verbanden tussen deze collecties te leggen die op geen enkele andere manier mogelijk zijn en ook niet op voorhand te voorspellen zijn. Er ontstaat een zogenaamd semantisch netwerk dat complex verspreide kennis toegankelijker maakt en dat gecombineerd kan worden met grafische overzichten of concrete 'kennis-snipjets'. Op basis van de informatie in het semantische netwerk kunnen zoekmachines betere en intelligentere resultaten geven en kan relevante informatie makkelijker aangeboden worden.

ECHOES 2.0

Empowering Communities with a Heritage Open Ecosystem

Erfgoed Leiden en Omstreken, Tresoar, Generalitat de Catalunya, Diputació de Barcelona en Consorci de Serveis Universitaris de Catalunya werken samen aan de creatie van een ecosysteem voor het delen van digitale erfgoedcollecties, toegepast op de wensen van gebruikers en erfgoedinstellingen. Een ecosysteem als onderdeel van een gedistribueerd netwerk van erfgoedcollecties en verrijkmogelijkheden.

Doel van het totale project is de creatie van een universeel toepasbare, modulaire, open source IT-infrastructuur, of liever gezegd: een ecosysteem voor het delen van digitale erfgoedcollecties. Het ecosysteem verbindt goede open source oplossingen, die al gecreëerd zijn voor een deel van de behoeften van gebruikers en erfgoedinstellingen zoals het eerder beschreven RedBot en voegt daar nieuwe functionaliteiten, die nu nog gemist worden, aan toe.

Het project draagt bij aan het creëren van open toegang met gebruikmaking van gedistribueerde netwerken en verrijkmogelijkheden. Het product wordt gerealiseerd in internationaal verband door regionaal, landelijk en internationaal acterende instellingen die in de diverse culturele en technische sectoren (archieef, bibliotheek, museum, universiteit) werkzaam zijn.

Waar mogelijk worden data omgezet in RDF-triples zodat zij gepubliceerd kunnen worden als Linked Open Data, dat is één van de basisprincipes van het project. Een ander basisprincipe is dat de modules gebruiken maken van standaard vocabulaires / trefwoordenlijsten via Open SKOS³⁶ en bijvoorbeeld DBpedia.

Het project draagt bij aan het creëren van open toegang met gebruikmaking van gedistribueerde netwerken en verrijkmogelijkheden. Verrijking van data wordt binnen dit ecosysteem ondersteund door de mogelijkheid (geo)informatie toe te voegen, informatie te verbinden en crowdsourcing.

Nationaal Archief

Het Nationaal Archief onderkent de volgende uitdagingen voor archiefinstellingen:

- Verbinden van de eigen archiefcollectie aan andere collecties
- Nadere (inhoudelijke) ontsluiting

Voor het aan elkaar verbinden van collecties is nu vaak afhankelijk van het gebruik van een zelfde beschrijvingsmethodiek. Dergelijke collecties kunnen veelal geharvest worden door aggregatoren als Archives Portal Europe, maar zijn alleen doorzoekbaar op velden uit de beschrijvingsmethodiek. Om te verbinden aan andersoortige collecties wordt er veelal een “kleinst gemene veelvoud” van beschrijvingen gehanteerd (zoals Dublin Core) die informatie plat slaat, dit om het probleem te slechten dat er met verschillende standaarden wordt gewerkt.

³⁶ <http://openskos.beeldengeluid.nl/>

Voor de nadere ontsluiting doemen al snel vragen op als: Welke termen gebruik je? Hoe kun je daarop zoeken? Is er een mini-schema nodig voor elk soort nadere toegang?

Het Nationaal Archief zet Linked Data nadrukkelijk in als toegangsinstrument, als oplossing voor de geschetste problemen. Doordat brokjes kennis granulair beschreven worden (in triples), zit je niet vast aan metadataschema's, je kunt data koppelen. Een deel van de data zal geautomatiseerd gekoppeld kunnen worden aan termenlijsten (snel, niet 100% correct), sommige data zal handmatig in het collectiebeheersysteem aan termenlijsten gekoppeld worden (niet snel, hoge kwaliteit).

De vraag die hierbij wel opkomt is: welke begrippen gebruiken we? Wie beslist dat "Rembrandt" en "Rembrandt van Rijn" dezelfde persoon zijn? Als je zoekt op "sport", hoe weet je dan dat "voetbal" ook relevant is? Hierbij wordt het MARA principe "liever verwijzen dan beschrijven" gehanteerd, oftewel: gebruik in de beschrijving liever concepten dan woorden. Achteraf beslissen dat twee begrippen verwijzen naar hetzelfde concept kan altijd, maar als je in je beschrijving direct al hetzelfde concept gebruikt scheelt dat werk achteraf.

Een andere vraag die beantwoordt moet worden: welke concepten hebben we in ieder geval nodig? De gebruiker zoekt vaak vanuit Wie, Wat, Waar, Wanneer. Er zijn dus concepten nodig voor Personen, Onderwerpen, Plaats/locatie, Periode/datum/tijdvakken/gebeurtenissen. Idealiter zijn er voor elk van deze concepten enkele voorkeurslijsten. Het Nationaal Archief maakt hiervoor een archief ontologie, dat onderdeel zal uitmaken van het termen netwerk van het NDE.

Dit netwerk van terminologiebronnen³⁷ ("Knowledge Graph voor Digitaal Erfgoed") is de verzameling van gelinkte terminologiebronnen die als referentie dient voor het maken van verwijzingen naar plaatsen, personen, onderwerpen, tijdsperiode, gebeurtenissen in de beschrijvingen van de collecties en objecten. Het doel hiervan is tweeledig: zorgen voor een hoogwaardig netwerk van referenties en het zo breed mogelijk implementeren hiervan. Uiteindelijk is dit een open netwerk waar iedereen onder bepaalde voorwaarden aan kan bijdragen. De basis van deze gemeenschappelijke voorziening wordt gevormd door de knooppunten en andere organisaties die gezaghebbende terminologiebronnen en andere referenties onderhouden en beschikbaar stellen bijv. ook WikiData, Getty, OCLC, enzovoort.

De Linked Data van het Nationaal Archief, deels gevoed vanuit het eigen collectiebeheersysteem, zal via een component beschikbaar komen voor gebruik in de nieuwe publiekswaasite van het Nationaal Archief. Naast het eigen gebruik zal de Linked Data tevens als Linked Open Data voor hergebruik beschikbaar wordt gesteld.

³⁷ <http://www.den.nl/pagina/522/bruikbaar/>

7. Visie van leveranciers op het Semantisch Web

Het gebruik van Linked Data staat niet op zich. Veelal zal er een relatie zijn met het collectie beheersysteem waar data in wordt beheerd en ook software die de data presenteert. Aan diverse leveranciers is gevraagd naar hun visie op het Semantisch Web. De ontvangen input is hier integraal opgenomen.

DEVENTit

Inleiding

Atlantis is een volledig webbased omgeving voor collectiebeheer en portaalbouw. Eén van de doelstellingen is dat de te beheren of gekoppelde objecten op allerlei manieren gecontroleerd beschikbaar kunnen worden gesteld voor gebruik door websites, portalen, collectiebeheersystemen, etc. Daarbij hoort ook het Semantisch Web.

Open Data

Atlantis heeft een dienst "Open Data". Daarmee kan een organisatie willekeurige selecties en combinaties van data beschikbaar stellen in willekeurige opmaak. Dit leidt tot een Open Data Portaal. Onder de mogelijk beschikbaar te stellen opmaken kan de informatie dus ook in formaten beschikbaar gesteld worden die het Semantisch Web voeden. Bijvoorbeeld in de vorm van RDF, OWL en/of XML opmaak.

Linked Open Data

Met Atlantis kunnen externe, gestandaardiseerde, thesauri gekoppeld worden met objecten in Atlantis. Hiermee worden objecten aan internationale, taalafhankelijke en sector overschrijdende terminologieën gekoppeld via de unieke identificaties van deze termen (URI's). Hiermee komt de informatie van objecten als Linked Open Data beschikbaar via het Open Data Portaal.

Semantisch Web

Het Semantisch Web gaat de komende jaren een steeds hogere vlucht nemen als omgeving voor ontsluiting van informatie. Atlantis biedt de noodzakelijke tools aan om de informatie aan dit Semantisch Web aan te kunnen bieden. Ontwikkelingen op dit vlak worden nauwlettend in de gaten gehouden. Nieuwe standaarden en methoden zullen dan ook steeds in Atlantis geadopteerd worden en zo beschikbaar komen voor organisatie die Atlantis toepassen voor hun collectiebeheer en portaalbouw.

Spinque

De principes van Linked Open Data verlagen de drempels voor het (her)-gebruik van gestructureerde informatie. Met Spinque wordt het gebruik van Linked Data ook een realiteit voor praktische toepassingen.

Een belangrijk aspect van Linked Data is dat bij het publiceren van de data al rekening wordt gehouden met de integratie met andere data. Denk hierbij aan het publiceren volgens standaard data representatie talen, het uniek identificeren van objecten, het modelleren met behulp van bestaande vocabulaires en het leggen van links naar andere databronnen. Het resultaat is een graaf waarin objecten met elkaar verbonden zijn via relaties (links).

Onderzoek naar het semantische web heeft de query taal SPARQL ontwikkeld om Linked Data bruikbaar te maken. Met deze query taal over grafen kunnen complexe vragen worden geformuleerd om data uit verschillende bronnen te combineren. In de praktijk blijkt echter dat voor veel toepassingen van Linked Data het gebruik van een dergelijke query taal alleen niet voldoende is; zoeken op tekst is vaak ook een vereiste. Idealiter worden hierbij de resultaten gerankt op relevantie. Deze ranking overstijgt de zoekfuncties van SPARQL die gebaseerd zijn op het letterlijke string-matchen. Wie weet immers exact de URL van een bepaalde persoon? Die zoeken we toch liever op via de naam. Ook het vinden van personen die veel weten over een onderwerp (expert finding) of het vinden van gerelateerde objecten (recommendation) vereist zoektechnologie. Applicaties van Linked Data maken daarom vaak gebruik van zowel een triple store als een zoekstelsel.

Het inrichten van een zoekstelsel over Linked Data zodat het effectief en efficiënt werkt is echter niet eenvoudig. Zoeksystemen zijn ontworpen voor het zoeken naar documenten (of platte records met metadata velden), waarin de redundantie van natuurlijke taal wordt benut om goede antwoorden te geven. Terwijl de kracht van Linked Data juist zit in de graaf structuur waarin objecten uit verschillende bronnen met elkaar zijn verbonden. Om bestaande zoeksystemen te gebruiken voor Linked Data wordt de data uit de graaf platgeslagen tot de documenten (records) waarop we willen zoeken. Door deze transformatie gaat de flexibiliteit en expressiviteit van het graaf model verloren.

Spinque is het innovatieve alternatief voor het beantwoorden van complexe zoekvragen over Linked Data. Spinque combineert de kracht van zoektechnologie met de flexibiliteit van het graaf model. Met Spinque Desk maken informatiespecialisten zelf een zoekmachine, door in een grafische omgeving zoekstrategieën te modelleren. Dezelfde techniek vormt de basis van CultuurLINK, een service die Spinque heeft gerealiseerd voor het Nederlands Instituut voor Beeld & Geluid.

Met Spinque Desk wordt het eenvoudig om informatie uit thesauri te gebruiken bij het zoeken naar records die beschreven zijn met de termen uit deze thesauri. Met hetzelfde systeem kan ook een zoekmachine worden gemodelleerd om experts te vinden over een onderwerpen op basis van documenten die ze hebben geschreven. Nog een ander voorbeeld is het gebruik van links om suggesties te geven over verschillende databronnen.

Met initiatieven als DBPedia, WikiData en schema.org ontvouwt Linked Data zich als de standaard voor het web van data. Spinque maakt het mogelijk om Linked Data op een effectieve en efficiënte manier te gebruiken voor praktische toepassingen.

8. Naar de toekomst

In veel initiatieven voor vernieuwing wordt de focus gelegd op Linked Data, dit vraagt om reflectie. Waarom is er behoefte aan Linked Data? Welk probleem lost de nieuwe technologie op of welke nieuwe mogelijkheden maakt het mogelijk? Wie vraagt erom? Wie zijn eigenlijk de gebruikers van Linked Data? In welke behoefte van gebruikers wordt er door Linked Data voorzien? Kan deze behoefte niet met bestaande technologie opgelost worden? Linked Data is een veelbelovende technologie, maar door het ‘hogere’ doel in zicht te houden voorkom je een *technology push* zonder gewenste resultaten.

Ervaring opdoen met een nieuwe technologie maakt een archiefinstelling mondiger richting leveranciers en maakt dat projectplannen met meer inzicht geschreven kunnen worden. Met name de impact van Linked Data op de eigen organisatie is hierbij een belangrijk punt. Het mantra “verbinden, verbinden, verbinden” heeft immers niet alleen met technologie te maken. Er is een belangrijke taak weggelegd voor archiefinstellingen bij het daadwerkelijk verbinden en kennis toevoegen aan bestaande data. Deels kan dit geautomatiseerd, deels zal dit ook handwerk zijn. Daar er bij archiefinstellingen veelal veel kennis over de eigen data is, ligt het voor de hand het semantiseren bij de bron te beleggen en niet bij aggregatoren waar minder kennis is. Waar het nu een taak is van archiefinstellingen om de archieven te beschrijven wordt het beschrijven van de data een belangrijke taak in de toekomst.

Wanneer er collecties verbonden zijn, er sprake is van een bijdrage aan het Semantisch Web, dan is er een goed fundament gelegd. Echter, dat is slechts een voorbereiding op de mogelijkheden van ontsluiting. Er zijn dan nog geen gebruikerstoepassingen en een gewone gebruiker maak je niet blij met triples in een triple store. Het visualiseren van de verbindingen, het gebruik maken van kennis in zoekmachines, het faciliteren van het vertellen van een verhaal, het opzetten van themawebsites en het analyseren van data, zijn vervolgstappen die niet vergeten moeten worden. Een mooie toekomst ligt dan voor ons.

9. Meer kennis

Door te lezen

- “Nederlandse Parels van Linked Data Toepassingen”, <http://www.geonovum.nl/sites/default/files/LinkedDataParels.pdf>, Platform Linked Data Nederland - Geonovum, 2015.
- “Linked Data for Libraries, Archives and Museums; How to clean, link and publish your metadata”, [http://book.freeyourmetadata.org/Seth van Hooland en Ruben Verborgh](http://book.freeyourmetadata.org/Seth%20van%20Hooland%20en%20Ruben%20Verborgh), 2014.
- “Linked data in beeld”, http://www.pilod.nl/w/images/3/32/Linked_Data_in_beeld_2014.pdf, Platform Linked Data Nederland, 2014.
- “Pilot Linked Open Data” Deel 1: Het Managementoverzicht (https://issuu.com/geonovum/docs/gnv_lod_boekwerk_01), Pilot Linked Open Data - Geonovum, 2013.
- “Pilot Linked Open Data” Deel 2: De Verdieping (https://issuu.com/geonovum/docs/gnv_lod_boekwerk_02), Pilot Linked Open Data - Geonovum, 2013.
- “Linked Data: Evolving the Web into a Global Data Space”, <http://linkeddatabook.com/editions/1.0/>, Tom Heath en Christian Bizer, 2011.

Door te bekijken

- Hoe de brandweer van Amsterdam-Amstelland Linked Data gebruikt, http://videlectures.net/iswc2012_van_leeuwen_emergency_response/
- De ervaringen van de Duitse uitgever Springer Nature, <http://www.slideshare.net/mpasin/linked-data-experiences-at-springer-nature>
- Hoe het Britse omroep BBC gebruikt maakt van Linked Data, <http://www.iskouk.org/sites/default/files/KentSlides2016-07-19.pdf>

Door te participeren

- Platform Linked Data Nederland, <http://www.pilod.nl/>
- Linked Data community, <http://linkeddata.org/>
- Kennisplatform toegang, <https://informatie2020.pleio.nl/groups/profile/32601572>