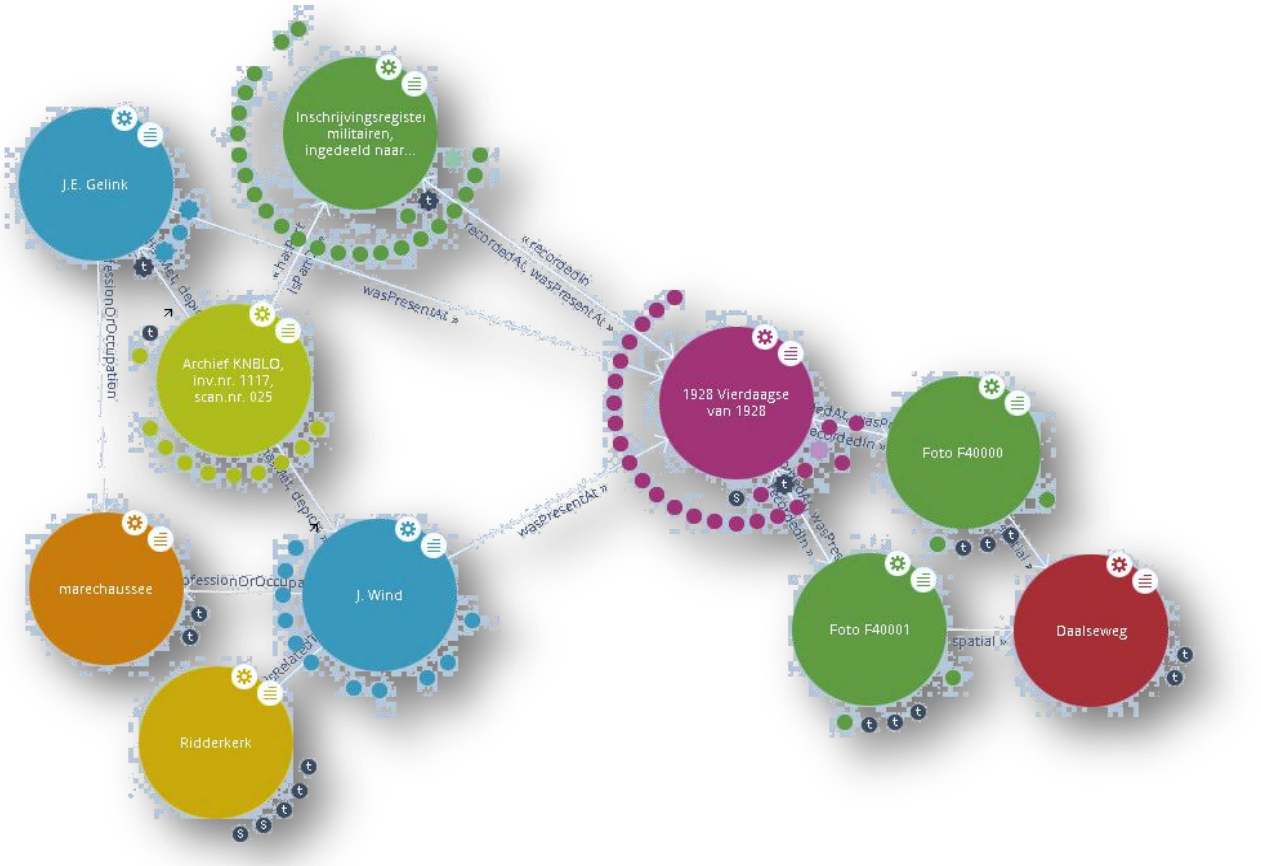
GEMEENTE Nijmegen

Stappenplan Linked Open Data voor Archieven



Regionaal Archief Nijmegen

16-11-2016

Colofon

Samenstelling

Dit stappenplan is het resultaat van de pilot in het kader van het project Stappenplan Linked Open Data ; Historische datasets vierdaagse van Nijmegen semantisch verbinden

Tekst

Lieke Verhelst (Linked Data Factory)
Renier van de Giessen (Regionaal Archief Nijmegen)
Henk Trapman (Regionaal Archief Nijmegen)
Anite op de Woerd (Regionaal Archief Nijmegen)

November 2016

Stappenplan Linked Open Data voor Archieven

Contents

Voorwoord

In 2016 is de Nijmeegse Vierdaagse voor de 100e keer georganiseerd. Het Regionaal Archief Nijmegen (RAN) heeft deze gelegenheid aangegrepen om d.m.v. een pilotproject de historische gegevens van dit grootste wandelevenement ter wereld als Linked Open Data te publiceren. Dit project *Stappenplan Linked Open Data ; Historische datasets vierdaagse semantisch verbinden* heeft subsidie ontvangen van Archief 2020 in het kader van de *doelstelling toegankelijkheid, deelproject semantiek, linked data en geodata*. Het stappenplan is een van de eindresultaten van het project.

Het Regionaal Archief beheert diverse archieven en collecties over de vierdaagse. Beschrijvingen van foto's en van namen van deelnemers stelde het RAN al beschikbaar via de eigen website. Daarnaast waren op de website van het Huis van de Nijmeegse Geschiedenis gegevens te vinden over gelopen afstanden, vertrekplaatsen en aantallen deelnemers en uitvallers. Deze losse datasets zonder onderlinge verbindingen en koppelingen zijn in de pilot gebruikt om Linked Open Data van te maken.

Hoofddoel van het project was het onderzoeken van de te nemen stappen om datasets aan elkaar te verbinden via Linked Open Data. Door middel van de pilot met data over de vierdaagse van Nijmegen zijn deze stappen onderzocht.

Andere doelen die hieruit voortkwamen waren:

- Kennis vergaren over LOD
- Linked Open Data maken van de vierdaagse data
- De vierdaagse datasets aan elkaar koppelen
- Community oprichten voor kennisdeling en een breed gedragen stappenplan
- Stappenplan opstellen
- Seminar voorbereiden

Onder leiding van Lieke Verhelst zijn de projectmedewerkers gestart met het maken van Linked Open Data. In een leer/werk omgeving is kennis opgedaan over Linked Open Data, zijn vragen beantwoord en zijn de praktische stappen gezet die nodig zijn om Linked Open Data te maken.

De stappen bestonden uit:

- Data opschonen met behulp van Open Refine
- Vocabulaires zoeken
- Data modelleren op basis van het Europeana Data Model (EDM)
- Data verrijken door middel van het koppelen met Dbpedia, Geonames,
- Data omzetten naar Linked Open Data, de zogenaamde triples
- Triple store aanschaffen en inrichten
- Triples opslaan in de triple store

Dit heeft geleid tot de volgende resultaten:

- Gekoppelde datasets
- Verrijkte datasets
- Triplestore met triples
- Stappenplan
- Kennis over LOD
- Deelname aan studiedag LOD

Voor de pilot zijn bestaande data gebruikt. Deze data komen uit systemen die niet zijn ingericht voor het omzetten van data tot LOD of voor het maken van koppelingen. Het schonen en verrijken van de data heeft daardoor veel extra tijd gekost, vooral omdat het grotendeels handmatig moest. Dit heeft er wel toe

geleid dat er is nagedacht over het structureren en standaardiseren van gegevens in het eigen archiefbeheersysteem, zodanig dat data op een goede manier verrijkt kunnen worden. Omdat dit automatisch verrijken via een archiefbeheersysteem nog niet bestond is het geen onderdeel geweest van de pilot en ook niet vermeld in het stappenplan.

Voor de duur van het project is een triple store ingericht waar de gekoppelde vierdaagsedata is te zien, te doorzoeken en te hergebruiken http://vierdaagselod.nl/lodview/Dataset.html
Om derden meer tijd te geven om deze LOD te gebruiken blijft de triplestore nog tot eind 2017 beschikbaar.
Het RAN heeft hierdoor ook ruimte om praktisch bezig te blijven met LOD maar ook om enkele belangrijke strategische vragen te beantwoorden.

- Wil het RAN alleen triples aanbieden via een triple store of ook nog een publieksvriendelijke website bouwen voor het zoeken en tonen van de data
- Wil het RAN triples plaatsen in een eigen triple store of bij/door een aggregator
- Wil het RAN zelf triples maken of alleen open data daarvoor beschikbaar stellen aan een aggregator voor de omzetting naar LOD

Uit de pilot is gebleken dat alles zelf doen een tijdrovende klus is. Het is dus de vraag of een archiefinstelling dit moet doen en hiervoor alle kennis en technische inrichting voor moet hebben.

Op basis van alle opgedane ervaringen in de pilot is het stappenplan opgesteld. Dit is als concept gepresenteerd op de studiemiddag Linked Data Archieven op 20 september 2016 in het Nationaal Archief te Den Haag. Reacties konden tot 15 oktober ingeleverd worden. Hier is niet veel gebruik van gemaakt. Tijdens de pilot en het ontwikkelen van het stappenplan bleek al dat het moeizaam was om andere archiefinstellingen, instanties en personen mee te krijgen om mee te denken. De remmende voorsprong heeft hier een rol gespeeld. Het Nationaal Archief heeft aangeboden om vanuit hun expertise over met name de architectuur het stappenplan verder aan te vullen.

Inleiding

Dit stappenplan is een handreiking voor archieven die willen starten met Linked (Open) Data. Hoewel het speciaal voor archieven is geschreven, is het mogelijk ook voor andere (erfgoed) instellingen van toepassing of nuttig. Het plan is met name gemaakt om organisaties een globale impressie te geven van de kosten en baten van het publiceren van Linked (Open) Data. Daarnaast geeft het plan een opsomming van de stappen die genomen moeten worden om tot de publicatie van Linked Data te komen. Dit is voornamelijk een verwijzing naar een reeds bestaande, technische beschrijving.

In dit stappenplan wordt gebruik gemaakt van een aantal termen. Onderaan dit stappenplan staat een verklarende woordenlijst.

De stappen

STAP 1: ORIENTATIE

"Waarom Linked Data?"

Als je met Linked Data wilt beginnen heb je wellicht via presentaties of publicaties vernomen van de mogelijkheden die Linked Data biedt. Om een realistische inschatting te kunnen maken van de totale kosten en baten van een Linked Data traject is het raadzaam om eerst diepere kennis te vergaren.

STAP 2: BEPAAL SCOPE

"Voor wie?"

Als je overtuigd bent van de meerwaarde van Linked Data voor je collectie, stel dan vast wie de doelgroep is. Dit kunnen externe partijen zijn, maar de Linked Data technologie kan ook worden toegepast binnen de organisatie of tussen een beperkte groep organisaties.

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

De vaststelling van de scope is voor een groot deel bepalend voor de kosten van het Linked Data publicatie traject.

STAP 3: ZOEK PARTNERS

"Linked Data maak je samen"

Het koppelen van Linked Datasets heeft pas echt meerwaarde als beide bronnen het RDF formaat hebben.

STAP 4: MAAK LINKED DATA

"Doen is leren."

Haal de juiste kennis in huis om het project mee te starten. Stel een goed team samen. In het uitgebreide stappenplan van het Platform Linked Data Nederland staat uitgelegd hoe je Linked Data maakt.

STAP 5: PUBLICEER LINKED DATA

“We zijn live!”

Maak bekend dat je Linked Data publiceert! Je wordt niet vanzelf gevonden…

STAP 6: ONDERHOUD LINKED DATA

“..en verder..”

Als blijkt dat je gepubliceerde Linked Data gegevens door de tijd heen veranderen is het noodzakelijk om een redactie-publicatie proces in te richten, vergelijkbaar met ieder ander redactie-publicatieproces. Wijzigingen in je Linked Dataset zul je moeten metadateren en verwijzingen naar eerdere versies is noodzakelijk.

Verdieping bij de stappen

In dit deel van het stappenplan wordt meer uitleg en achtergrondinformatie gegeven bij de verschillende stappen.

STAP 1: ORIENTATIE

"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. Het lezen van publicaties of bijwonen van presentaties is niet voldoende om de nuances in de belofte van Linked Data te ontdekken. Het is raadzaam om een cursus "Inleiding tot het Semantisch Web en Linked Data" van Go Opleidingen of Linked Data Factory te volgen, zodat je kritische vragen kunnen worden beantwoord door iemand uit de praktijk. Je medecursisten stellen bovendien vaak andere vragen zodat je elkaars kennisbasis vergroot.

Zorg in ieder geval dat je goed het verschil kent tussen Linked Data, Linked metadata (SKOS) en het linken van documenten op het web.

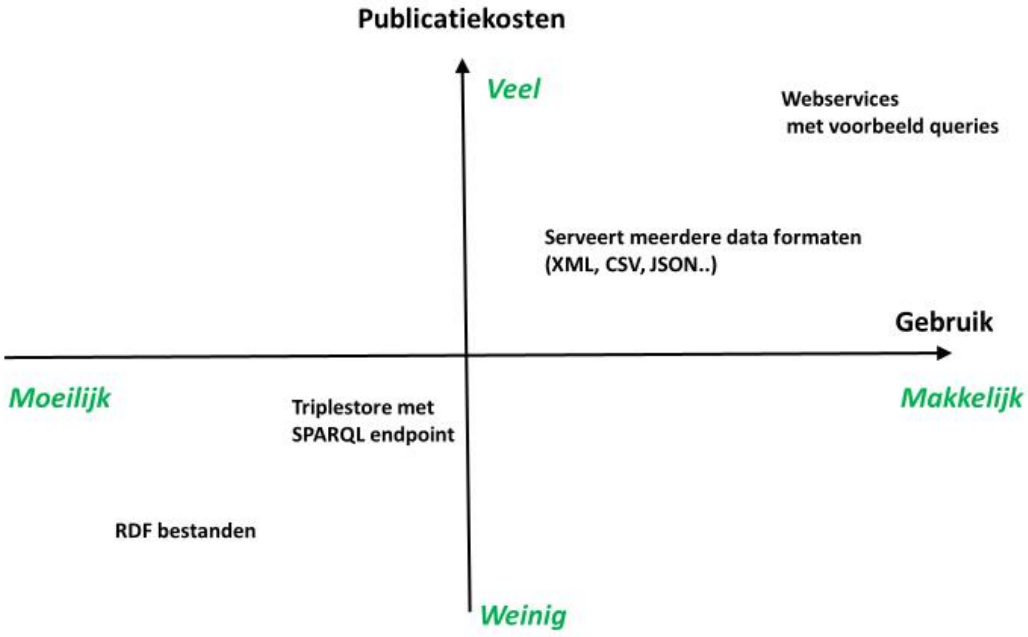
STAP 2: BEPAAL SCOPE

"Voor wie?"

Als je overtuigd bent van de meerwaarde van Linked Data voor je collectie stel dan vast wie de doelgroep is. Probeer hierbij specifiek te zijn, en te anticiperen hoe deze doelgroep jouw 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.

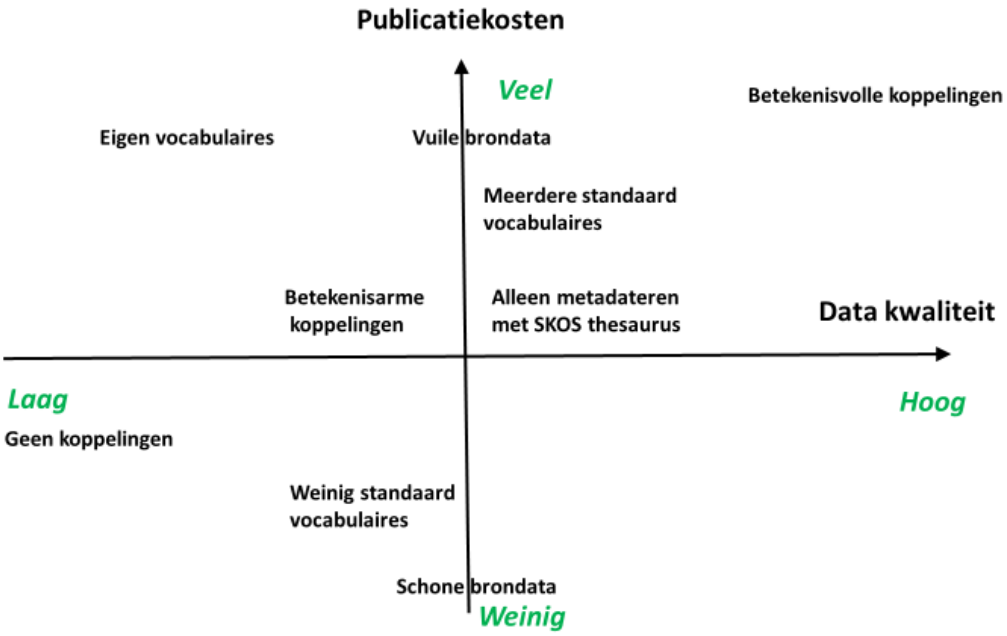
Bepaal ook of je je gegevens alleen wilt metadateren met een Linked Data thesaurus, of dat je alle gegevens in het Linked Data formaat wilt omzetten. Zie ook stap 2 onder het kopje "Vocabulaires".

Een algemene regel is: hoe breder de doelgroep, hoe hoger de totale publicatiekosten. (Met publicatiekosten worden alle kosten bedoeld die gemaakt worden bij het proces van het transformeren van bestaande data naar gepubliceerde Linked Data.) 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. Zie onderstaande afbeelding.



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 linker onder diensten zijn een voorwaarde voor de rechter boven diensten).

Er geldt verder ook: 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. Zie voor de technische stappen stap 4 van dit stappenplan, en voor de keuzes die je daarvoor vooraf moet maken onderstaande afbeelding.



De onderwerpen in deze afbeelding worden hieronder apart behandeld.

Schone/vuile brondata: 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.

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 Vocabularies

Veel erfgoedinstellingen hebben hun collecties al via een thesaurus-in-Linked-Data-formaat (=SKOS) gemetadateerd. Het resultaat hiervan is dat de metadata 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 GTAA van het Nederlands Instituut voor Beeld en Geluid en de erfgoedthesaurus van de Rijksdienst voor het Cultureel Erfgoed.

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-star 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 research 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.

STAP 3: ZOEK PARTNERS

“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 (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 uit stap 2. 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.

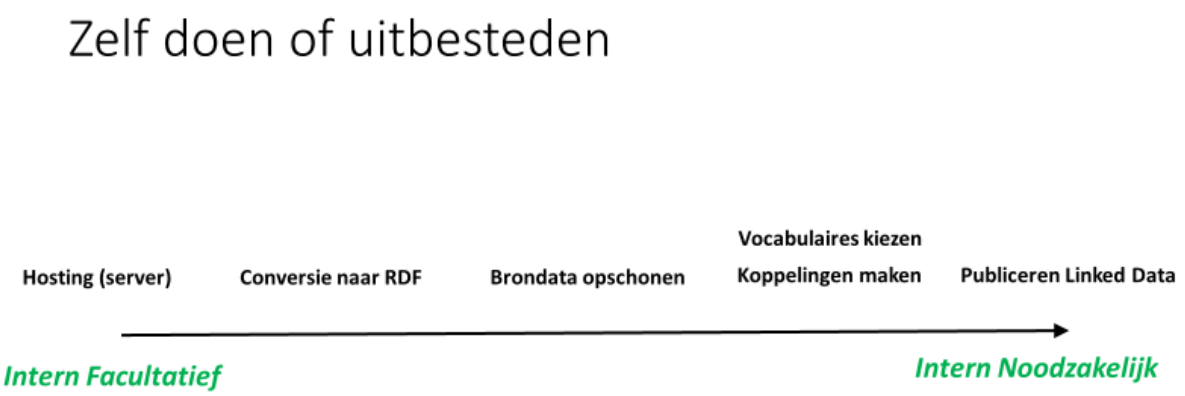
Als je hebt bepaald wie je doelgroep is, dan kun je daar waarschijnlijk vast contact mee maken. Kijk ook wat de andere partij voor jou kan betekenen. Hebben ze een dataset die voor jou interessant is om aan te koppelen?

STAP 4: MAAK LINKED DATA

"Doen is leren."

Bij het samenstellen van het Linked Data projectteam is het raadzaam verschillende deskundigheid op te nemen. In het team zitten uiteraard mensen met Linked Data kennis en kennis van webtechnologie, 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. 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 voor archieven het Europeana Datamodel en de Dublin Core vocabulaires . Gebruik voor het metadateren de thesauri van Beeld en Geluid of het RCE (zie stap 2, het kopje vocabulaires).

Niet alle dingen hoef je zelf te doen, maar sommige stappen moet je liever niet uitbesteden. Zie onderstaande afbeelding.



Een toelichting hierbij: 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 subdomein "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.

STAP 5: PUBLICEER LINKED DATA

"We zijn live!"

Je kunt om de vindbaarheid te vergroten je dataset publiceren in een Open Data catalogus zoals data.overheid.nl of de datahub. Hierin kun je aangeven dat je dataset het Linked Data formaat heeft. Er bestaat een speciaal vocabulaire dat bedoeld is om publicatie in een dergelijke catalogus te automatiseren. Dit is het DCAT vocabulaire.

STAP 6: ONDERHOUD LINKED DATA

"..en verder.."

Als je gepubliceerde Linked Data door de tijd heen veranderd moet worden omdat het niet meer actueel is, of uitgebreid moet worden, 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 zijn. Daarom besluiten de meeste organisaties om te beginnen met het publiceren van een Linked Data bestand waarvan ze weten dat het door de tijd heen niet zal veranderen.

Verklarende woordenlijst

Triplestore = software voor opslag van Linked Data

(zie ook: https://en.wikipedia.org/wiki/Triplestore)

RDF = manier om Linked Data op te slaan (technisch)

(zie ook: https://en.wikipedia.org/wiki/Resource_Description_Framework)

SPARQL Endpoint = open toegangspoort tot de Linked Data (technisch)

Vocabulaire = ander woord voor data model. Hierin wordt de betekenis van de gegevens vastgelegd.

Webservice = dienst die op bepaalde voorwaarden data serveert (technisch)

(zie ook: https://en.wikipedia.org/wiki/Web_service)

XML = uitwisselingsformaat voor data

CSV = opslagformaat voor data, te openen in spreadsheet software

JSON = uitwisselingsformaat voor data