



ADVIESRAPPORT

SEMANTIEK EN DATAMODEL REDBOT





Inhoud

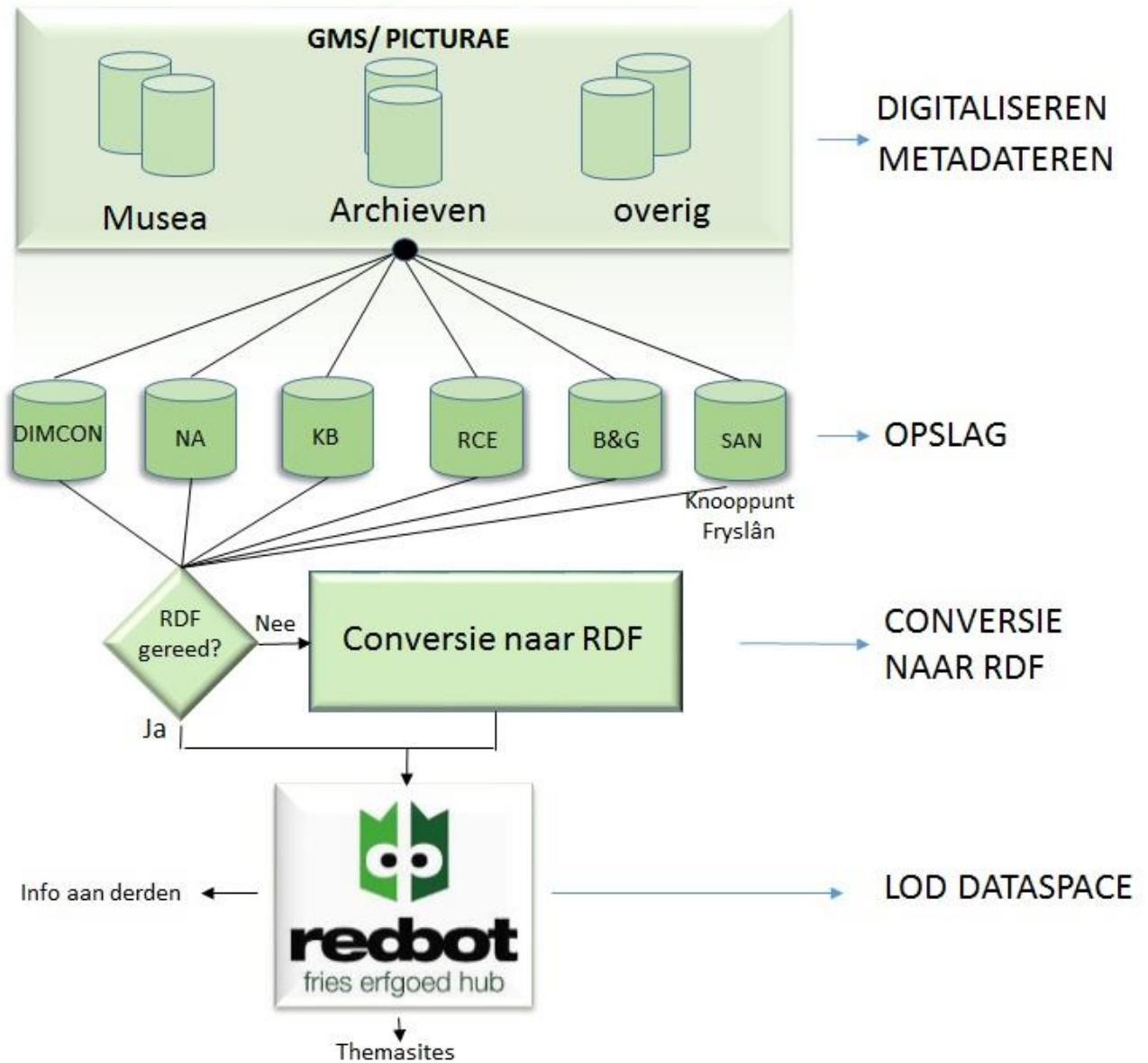
1 Inleiding	3
2 Managementsamenvatting	4
2.1 Adviezen	5
3 RDF Data in RedBot	7
3.1 Flexibiliteit	7
3.2 Semantische integratie	7
3.3 Optimaal gebruik van het Linked Open Data (LOD) netwerk.	8
3.3.1 Hergebruik van data	8
3.3.2 Standaardisatie met ontologie.....	8
3.3.3 Standaardisatie met 'Authority Data'.....	8
3.4 Het netwerk effect.....	9
4 RedBot als semantische hub.....	9
4.1 Participatie in het netwerk.....	10
4.2 Standaardisatie.....	11
4.3 Ontologie	11
4.3.1 Dublin Core Terms.....	11
4.3.2 Schema.org	12
4.3.3 Andere semantische modellen.....	12
4.3.4 Authority Data.....	12
5 RedBot als service hub: gedeelde technologie	17
5.1 Standaardisatie.....	17
5.2 RDF mapping datasets.....	17
5.3 Linken van datasets	17
5.4 Persistente Identifiers.....	18
5.4.1 Toelichting Persistente Identifiers	18
5.4.2 Rol van Redbot	18
6 Advies	19
6.1 Semantiek.....	19
6.2 Datamodel	19
6.3 RedBot Serviceorganisatie.....	20
7 Samenstelling werkgroep RedBot	21

1 Inleiding



Aanvullend aan het document 'Inpassing Friese Erfgoedhub in Frysk Kultuerhûs (RedBot) dat gaat over de infrastructuur en data, wordt in dit document het datamodel en de 'semantische laag' in de beoogde infrastructuur van RedBot nader toegelicht en wordt concrete invulling gegeven aan de rol van RedBot in het Linked Open Datanetwerk.

Het traject...



2 Managementsamenvatting

Onder de semantische laag wordt verstaan het netwerk van databronnen dat ontstaat door informatie uit de ene bron te relateren, of linken, aan informatie uit een andere bron. Hierbij is het nodig dat de verschillende bronnen beschikken over gedeelde



techniek en semantiek waarmee op uniforme wijze informatie wordt uitgewisseld. De semantische laag in RedBot heeft dus als doel het organiseren en realiseren van een Linked Open Data (LOD) netwerk met informatie over gedigitaliseerd Fries Cultureel Erfgoed.

RedBot, als mogelijke serviceorganisatie, kan een waardevolle rol spelen als facilitator bij de ontwikkeling van dit netwerk door ondersteuning en services te bieden op het gebied van conversie van data naar Resource Description Framework (RDF) en publicatie van die data op het Internet.

RedBot kan een rol spelen bij het standaardiseren van metadata in het Fries Cultureel Erfgoed domein en zo integratie van informatie van verschillende partijen in het domein faciliteren. Daarnaast kan RedBot gespecialiseerde services aanbieden die partijen nodig hebben om deel te kunnen nemen aan het LOD netwerk van Fries en Nederlands Cultureel Erfgoed. Dat geldt op lokaal, nationaal en internationaal niveau. Deze inspanningen maken het mogelijk dat er een verrijkt en uitbreidbaar datanetwerk beschikbaar komt.

2.1 Adviezen

Semantiek

Korte toelichting

Via RedBot worden collecties van de collectiehouders met elkaar en met de informatie infrastructuur van het internet verbonden.

(conform laag 'bruikbaar' in het drielagenmodel van de Nationale Strategie Digitaal Erfgoed).

Door het leggen van verbindingen vergroot dit de eenduidigheid van de data uit de afzonderlijke collecties en is als tweede stap deze data te verrijken met content uit de grote instellingen in het landschap van digitaal erfgoed Nederland en bronnen van de internet informatie infrastructuur.

Advies semantiek:

- Gebruik RedBot als verbinder van de Friese (erfgoed) collecties conform het drielagenmodel van de Nationale Strategie Digitaal erfgoed;
- Gebruik Linked Data techniek via standaard RDF;
- Registreer de gedigitaliseerde materialen, waar mogelijk, op landelijk of provinciaal niveau;
- Laat het converteren van de data naar (RDF) triples binnen RedBot te verzorgen.



Datamodel

Korte toelichting

Om de diverse data met elkaar te verbinden en betekenis te kunnen geven, adviseert RedBot een datamodel dat uitgaat van vijf basisuitgangspunten: wie, wat, waar, wanneer en hoe.

De eerste vier (wie, wat, waar, wanneer) zijn collectie-inhoudelijk, de vijfde (hoe) betreft de wijze waarop (digitale) objecten tot stand zijn gekomen. RedBot realiseert deze modellering via standaardisering en het hergebruik van gestandaardiseerde, breed gebruikte metadata modellen of ontologie. Deze efficiëntie wordt het meest gerealiseerd door per project bij digitalisering en invoeren van beschrijvende metadata (beschrijvingen) alvast voor te sorteren op gebruik van de standaarden en de authority files zoals DBpedia. Als collectiehouders bij de invoer al rekening kunnen houden met invoer volgens standaarden en volgens links naar authority files, dan verloopt de onderlinge uitwisselbaarheid en uitlijning met bv. een DBpedia veel soepeler hetgeen resulteert in nauwkeurige integrale collectieontsluiting voor de eindgebruiker.

Advies datamodel:

- De werkgroep adviseert om het toekennen van metadata te systematiseren en onderling afspraken hierover te maken tussen RedBot en de collectiehouders.
- De voorkeur gaat uit naar metadatering volgens de 'Dublin Core Terms' standaarden en het datamodel vooraf met de collectiehouder af te stemmen (zie pa. 4.4)
Voor dit vooraf of tijdig afstemmen van metadatering biedt RedBot diensten aan als onderdeel van RedBot Service Organisatie (zie pa. 6.3).

RedBot Serviceorganisatie

Korte toelichting

Fries digitaal erfgoed in een informatie infrastructuur via een verbindingslaag (RedBot) is meer dan het implementeren en toepassen van techniek. Een Serviceorganisatie kan collectiehouders ondersteuning bieden in het integraal ontsluiten van Fries digitaal erfgoed via het knooppunt RedBot. RedBot kan een waardevolle rol spelen in deze ondersteuning, zowel voor de collectiehouders voor goede integratie van de collecties naar het Friese knooppunt, als ook voor kwaliteit van de eigen metadata.

Advies serviceorganisatie:

- De werkgroep adviseert RedBot de rol op zich te laten nemen als Service Organisatie.
RedBot beschikt over kennis op het gebied van digitaal erfgoed en heeft de beschikking over een groot netwerk voor digitaal erfgoed. Technische diensten die RedBot kan leveren, verhogen de kwaliteit van de collecties van de collectiehouders (en daarmee van Fries digitaal erfgoed) (Standaardisering, pa. 5.1) en maken dit digitale erfgoed duurzaam toegankelijk (Persistent identifiers pa. 5.4). Het helpt de collectiehouders de data volgens de standaarden en de Datamodellering onder te brengen in de vijf grote domeinen.
- De werkgroep adviseert vorm en inhoud van de RedBot serviceorganisatie te laten beschrijven en ter goedkeuring voor te leggen aan de stuurgroep RedBot.



3 RDF Data in RedBot

Geadviseerd is de data in RedBot te beschrijven middels Resource Description Framework (RDF). De overwegingen die ten grondslag liggen aan die keuze zijn de volgende en worden hieronder nader toegelicht:

1. Flexibiliteit bij het opbouwen van gekoppelde dataverzamelingen/-bronnen.
2. De mogelijkheid van semantische integratie van diverse databronnen/-verzamelingen.
3. Optimaal gebruik maken van het Linked Open Data netwerk of het Semantisch Web.
4. Het netwerk effect.

3.1 Flexibiliteit

De doelstelling van RedBot om diverse databronnen en -verzamelingen te koppelen, te verrijken, en te ontsluiten vraagt om flexibiliteit in de te gebruiken methodiek vanwege onvoorziene diversiteit in databronnen en -verzamelingen. Het is noodzakelijk dat de modellering van data meegroeit met het semantische bereik van de gekoppelde data. Daarvoor is het noodzakelijk dat de metadata modellering los van data in het systeem kan worden aangepast en uitgebreid. De structurering van data als (RDF) triples biedt die flexibiliteit;

Triples bevatten uitspraken over de kenmerken van de bronnen in de vorm van een driedelige subject-predicaat-object-structuur. Het metadata model kan meegroeien met de diversiteit aan informatie die gaandeweg ontstaat naarmate het datanetwerk groeit.

3.2 Semantische integratie

Het koppelen van datasets of -bronnen gebeurt op semantisch niveau, bijvoorbeeld door Personen of Plaatsen in verschillende datasets aan elkaar te relateren. Die relatie, of link, kan allerlei betekenissen hebben. Zo is de relatie van semantische gelijkwaardigheid een veelgebruikte link tussen datasets waarmee wordt beschreven dat een Plaats in dataset A dezelfde is als een Plaats in dataset B. Door de definitie van semantische gelijkwaardigheid wordt vastgesteld dat *alle* eigenschappen die zijn toegekend aan de Plaats in dataset A ook van toepassing zijn op dezelfde Plaats in dataset B, en vice versa. Er wordt dus bereikt dat de informatie in beide datasets over dezelfde Plaats elkaar aanvult en de totale informatie die kan worden aangeboden vanuit het systeem over de betreffende Plaats toeneemt.

Bovenstaand voorbeeld van verrijking van data door het definiëren van links speelt zich af op data niveau, namelijk informatie over individuen. Ook op metadata niveau kunnen semantische relaties worden beschreven die de onderlinge verhouding tussen verschillende metadata modellen weergeven. Op een hoger abstractieniveau wordt hiermee hetzelfde bereikt als in bovenstaand voorbeeld, namelijk de integratie van data uit verschillende bronnen die beschreven zijn volgens verschillende metadata modellen.

Classificatie van informatie is een belangrijk aspect bij de semantische integratie van datasets of -bronnen. De beschrijving van objecten, relaties tussen objecten, en eigenschappen van objecten wordt gebaseerd op een classificatie van begrippen binnen de datasets. Deze classificatie is een opsomming van klassen binnen het metadata model en beschrijft de semantiek van objecten, bijvoorbeeld de veelgebruikte klasse Persoon of Plaats.



Stel nu dat in dataset A de klasse Persoon wordt gebruikt en in dataset B de klasse Kunstschilder, dan is het nodig voor de integratie van data om op metadata niveau aan te geven dat een Kunstschilder een subklasse is van Persoon. Hierdoor wordt de informatie in dataset A vergelijkbaar met dataset B want een Kunstschilder is dus ook een Persoon en daarom zijn alle eigenschappen van Personen beschreven in dataset A van toepassing op dezelfde Personen in dataset B. Let op, het omgekeerde is niet correct omdat een Kunstschilder meer specifieke eigenschappen heeft dan een Persoon (maar wel alle eigenschappen van Persoon!).

Het belang van classificatie is gelegen in het realiseren van een gemeenschappelijk begrippenkader. Door data uit verschillende datasets te conformeren aan een gemeenschappelijk begrippenkader is een eerste stap in de integratie van datasets gemaakt. Middels classificatie is het mogelijk om specialistische informatie te koppelen aan meer generieke informatie zoals in het bovenstaande voorbeeld beschreven.

3.3 Optimaal gebruik van het Linked Open Data (LOD) netwerk.

3.3.1 Hergebruik van data

Het LOD netwerk is een uitdijende verzameling van onderling gekoppelde RDF databronnen. Het LOD netwerk bevat een schat aan informatie die beschikbaar is voor hergebruik. Hergebruik van informatie is belangrijk omdat standaard terminologie het collectieve begrip van zaken verbetert. Daardoor wordt het uitwisselen van informatie eenvoudiger. In het LOD netwerk is een schat aan data en metadata beschikbaar voor gebruik ter beschrijving *en* verrijking van 'eigen' data.

3.3.2 Standaardisatie met ontologie

Bronnen als Dublin Core Metadata Initiative, Schema.org, DBpedia.org en anderen bieden breed gebruikte ontologie als metadata model ter beschrijving van data. Door van meet af aan gebruik te maken van dergelijke LOD bronnen (i.e. ontologiën) wordt een niveau van standaardisatie gegarandeerd over verschillende datasets of -bronnen heen. Dergelijke standaardisatie vereenvoudigt het aansluiten van databronnen in het LOD netwerk vanwege de *á priori* gedeelde semantiek.

3.3.3 Standaardisatie met 'Authority Data'

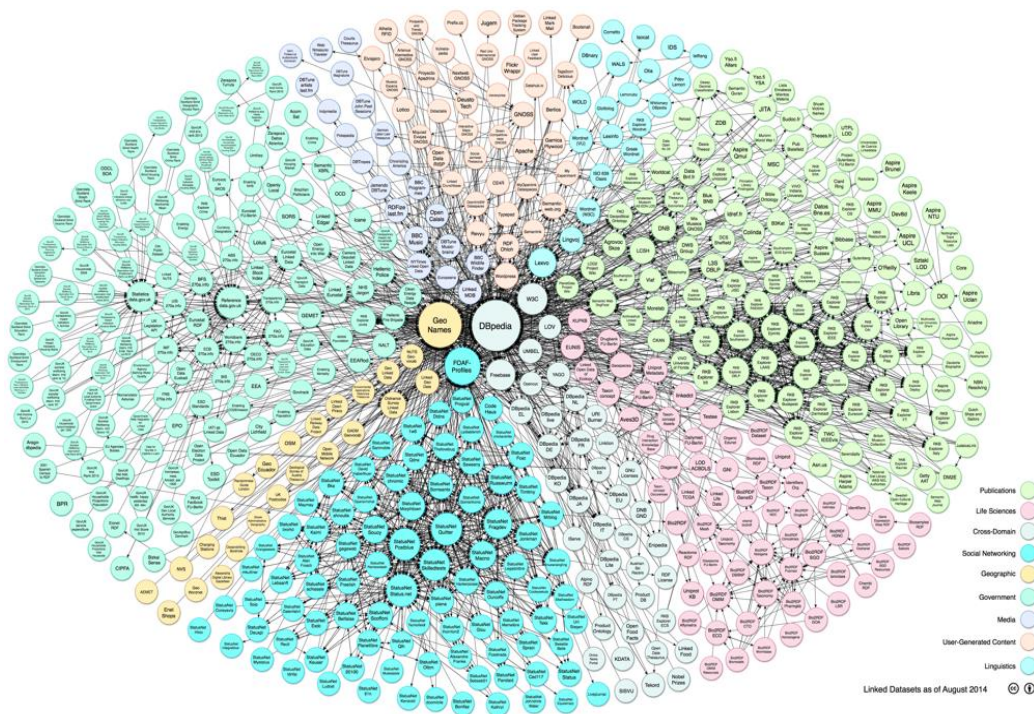
Een LOD bron zoals DBpedia (Wikipedia) biedt grote hoeveelheden gedetailleerde informatie over onderwerpen zoals Plaatsen, Personen en andere Dingen. Een link tussen een Plaats in dataset A en een Plaats in DBpedia biedt uitgebreide aanvullende informatie over de Plaats in dataset A. Bovendien is via DBpedia de uitgebreide thesaurus beschikbaar zoals die in Wikipedia wordt gehanteerd. Via deze thesaurus wordt de Plaats in dataset A gerelateerd aan andere Plaatsen of Personen die in DBpedia worden beschreven. Een link met DBpedia biedt verrijking aan een dataset door de koppeling van concepten binnen de dataset met concepten in het LOD netwerk. DBpedia is het centrale knooppunt in het LOD netwerk, vele datasets linken van en naar DBpedia waardoor data in DBpedia een autoriteit in het semantisch netwerk is geworden. Dit is niet zo vreemd want de bron van DBpedia is Wikipedia. Het gebruik van data van een autoriteit ('authority data') heeft voordelen vanwege standaardisatie. Stel de Plaats in dataset A en de Plaats in dataset B zijn beschreven in termen van DBpedia dan is het niet meer nodig om de link tussen beide datasets te zoeken want deze is al duidelijk door het gebruik van 'authority data' uit DBpedia. Kortom, door binnen verschillende datasets gebruik te maken van centrale beschrijving van begrippen wordt standaardisatie bereikt waarmee de onderlinge koppeling van die datasets sterk wordt vereenvoudigd.



3.4 Het netwerk effect

Door RedBot als centrale hub te positioneren in een netwerk van informatie over Fries Cultureel Erfgoed worden de voorwaarden gecreëerd die nodig zijn om het netwerk verder efficiënt uit te breiden met diverse Friese (Erfgoed) bronnen. Er zal gebruik worden gemaakt van gangbare technologie in het LOD netwerk waarmee externe partijen eenvoudig kunnen aansluiten. In feite faciliteert RedBot de participatie van Friese (Erfgoed) Data in het uitdijende LOD netwerk¹.

Het netwerkeffect is de meerwaarde die ontstaat bij de toename van het aantal deelnemers aan een netwerk. RedBot als facilitator van het Friese (Erfgoed) Netwerk vindt hierin haar missie en haar doel; de missie is zo veel mogelijk participanten aanhaken aan het Friese Erfgoed Netwerk, en het doel is de optimale faciliteiten te verschaffen waarmee partijen eenvoudig en doeltreffend kunnen aansluiten bij het netwerk.



Afbeelding 1: De LOD Cloud

4 RedBot als semantische hub

Bovenstaande overwegingen bieden concrete aanknopingspunten voor de realisatie van RedBot als facilitator in het netwerk van Fries Cultureel Erfgoed. RedBot biedt de nodige faciliteiten aan partijen aan om deel te kunnen nemen aan het LOD netwerk. Het netwerk wordt getypeerd door het gebruik van RDF/OWL waarmee data en de semantiek ervan beschreven, uitgewisseld, en gerelateerd kan worden.

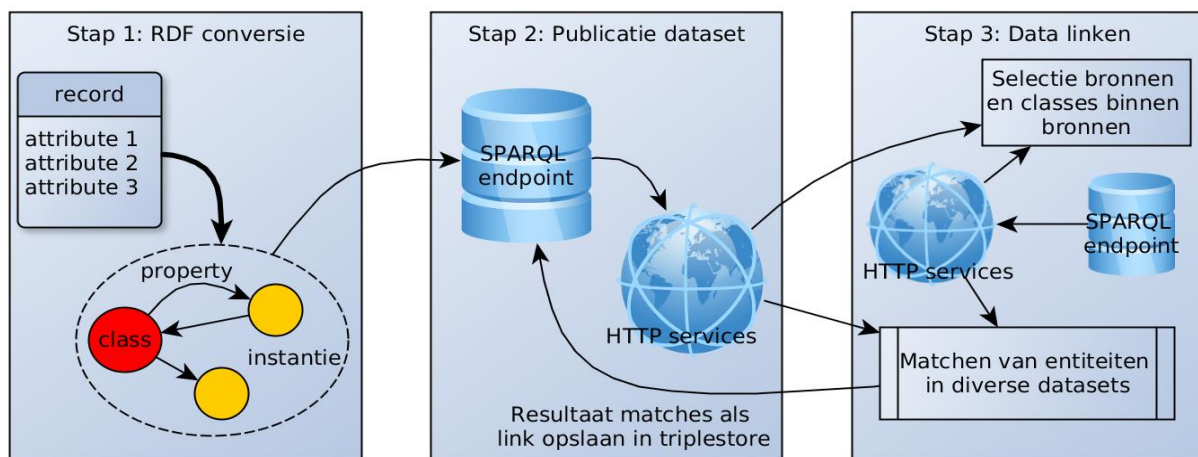
¹<http://lod-cloud.net/>



4.1 Participatie in het netwerk

Partijen kunnen op 2 manieren participeren aan het netwerk: door de realisatie van een eigen hub, of door data op een centrale hub te publiceren. De insteek is, daar waar partijen niet kunnen niet deelnemen aan een landelijk domein gaat RedBot fungeert als hoogste provinciaal domein. Het verschil tussen beide vormen is het wel of niet ontwikkelen van de benodigde services in eigen beheer. Voor meer duidelijkheid waar het hier om gaat worden de volgende stappen in het publiceren van een RDF dataset op een rij gezet:

1. Realisatie van een RDF representatie van de te publiceren (bestaande) datasets. Deze dataset moet uiteindelijk gerelateerd worden aan andere datasets in het LOD netwerk. Dat kan in eerste instantie door gebruik te maken van gedeelde ontologie en 'authority data'.
2. Publicatie van de dataset middels benodigde HTTP services, waaronder een SPARQL endpoint of andere methodiek waarbij 'resources' uit de dataset rechtstreeks toegankelijk zijn via hun IRI², bijvoorbeeld met een webbrowser.
3. Linken van de dataset aan andere databronnen door gelijke instanties van concepten met elkaar te verbinden en zo wederzijds aan te vullen, of te verrijken. Deze links vormen de vertakkingen in het LOD netwerk. Hier geldt de dynamiek: hoe fijnmaziger deze vertakking ontwikkelt, hoe rijker en beter de informatie in het netwerk wordt.



Een eigen hub realiseren heeft betrekking op de realisatie van de HTTP services en het SPARQL endpoint in stap 2. Elke partij die wil aansluiten bij het Friese Erfgoed Netwerk kan dat zelfstandig doen met services waar men al over beschikt of die men zelf ontwikkelt. Op deze wijze bestaat de mogelijkheid om tegen beperkte kosten zo veel mogelijk deelnemers aan het netwerk te faciliteren.

De realisatie van de benodigde HTTP en SPARQL services is specialistisch werk en partijen kunnen om die reden kiezen om gebruik te maken van centraal aangeboden HTTP en SPARQL services op het RedBot platform. Hierdoor verandert het hierboven beschreven proces niet inhoudelijk maar wordt de realisatie van de benodigde services uitbesteed aan een externe partij. RedBot kan door het inrichten van een efficiënt 'backoffice proces' schaalvoordelen behalen bij het aanbieden van centrale HTTP en SPARQL services.

²https://nl.wikipedia.org/wiki/Internationalized_resource_identifier



4.2 Standaardisatie

De ontwikkeling van het LOD netwerk is mogelijk geworden door het gebruik van gedeelde (metadata) modellen voor de representatie van informatie. Een metadata model beschrijft de semantiek van de gegevens in een dataset en hoe meer de semantiek in verschillende datasets overeenkomt hoe beter deze zijn te relateren, of linken. Het gemeenschappelijk gebruik van steeds beter ontwikkelde ontologiën in het LOD netwerk zorgt voor steeds beter te relateren informatie in verschillende datasets.

Het is van belang binnen het kader van RedBot een gemeenschappelijk begrippenkader vast te stellen *en* te ontwikkelen ter bevordering van participatie van bronnen met datasets in het Friese Erfgoed Netwerk. Wanneer alle betrokken partijen zich conformeren aan een generieke representatie van metadata dan wordt de ontwikkeling van het netwerk sterk bevorderd.

Het vaststellen van een gemeenschappelijk begrippenkader moet steunen op 'authority data' uit het LOD netwerk, ofwel gebaseerd worden op veel en breed gebruikte ontologie.

4.3 Ontologie

RedBot gaat in het datamodel in eerste instantie uit van een modellering die antwoord geeft op vijf basisuitgangspunten: wie, wat, waar, wanneer en hoe. RedBot realiseert deze modellering via standaardisering en het hergebruik van gestandaardiseerde, breed gebruikte metadata modellen of ontologie.

Het standaardiseren van een formeel omschreven begrippenkader gebeurt in RDF/OWL middels een gedefinieerde ontologie. Er zijn inmiddels vele domeinen beschreven met ontologie en hergebruik van dergelijke ontologie levert direct meerwaarde op, zoals hierboven beschreven is.

De keuze voor te gebruiken ontologie is afhankelijk voor de context waarin die gebruikt zal worden, 'fit-for-use', ofwel hoe goed past het begrippenkader zoals in de ontologie beschreven is bij het gebruik van data binnen RedBot.

Binnen de context van RedBot wordt vooral data over Fries cultureel erfgoed verzameld en verrijkt. Het doel is dergelijke data te publiceren voor een breed publiek en diverse toepassingen. Deels is inmiddels duidelijk welke relevante datasets worden opgenomen in RedBot en deels worden datasets nog gecreëerd door het digitaliseren van erfgoed objecten. Bestaande datasets³ bieden inzicht in het gebruik van metadata standaarden binnen de sector en het eventuele gebruik van 'authority data'.

4.3.1 Dublin Core Terms

Binnen bestaande datasets wordt veelvuldig gebruik gemaakt van Dublin Core⁴ gestandaardiseerde metadata. Vanuit overheidswege wordt Dublin Core standaardisatie ook geadviseerd in de vorm van OWMS⁵. Het gebruik van Dublin Core Terms binnen RedBot ligt daarom voor de hand. Dublin Core Terms voorziet echter niet in een passende classificatie voor het typeren van de dingen die worden beschreven en kent bovendien beperkte semantiek. Er is aanvullende ontologie nodig waarin gedetailleerde classificatie van concepten wordt beschreven die voldoende representatief zijn voor de

³<https://docs.google.com/spreadsheets/d/1PDRRhyTiODncDH-TxnrQeXRgUSsNptrtOaQhalhoJw/edit?usp=sharing>

⁴<http://dublincore.org/>

⁵<http://standaarden.overheid.nl/owms/terms>



datasets die binnen het netwerk beschreven moeten worden. Bovendien is het nodig dat die classificatie dusdanig ontwikkeld en georganiseerd is dat typering herleidbaar zijn tot algemene termen zodat brede vergelijking of doorzoeking wordt gefaciliteerd.

4.3.2 Schema.org

Schema.org biedt een goed ontwikkelde ontologie die geschikt is voor algemene classificatie van concepten. De ontologie van Schema.org wordt door veel partijen in het LOD netwerk gebruikt waaronder Google die o.a. ervan gebruik maakt om zoekresultaten te ordenen. Data die wordt aangeboden voorzien van Schema.org metadata formaat wordt optimaal meegenomen in de rangschikking van zoekresultaten. Schema.org heeft een algemene insteek, net als de DBpedia ontologie, maar kent een goed gestructureerde en behoorlijk volledige specificatie van algemene concepten. Schema.org is een geschikte algemene ontologie voor RedBot om te gebruiken als metadata model voor te verbinden datasets.

De generieke insteek van het metadata model zorgt voor een eenvoudige voorstelling van zaken. Zo blijft het eenvoudig in eerste instantie om partijen te betrekken bij het publiceren en relateren van RDF datasets. Binnen het domein van Cultureel Erfgoed bestaan echter al een praktijk van het gebruik van metadata modellen waarin meestal een meer complexe representatie van informatie wordt gegeven dan wenselijk is voor een eenvoudige benadering. Deze complexiteit komt meestal voort uit de problematiek van het catalogiseren van het aggregeren van informatie. Eindgebruikers zijn praktisch nimmer geïnteresseerd in dergelijke complexe informatie die vooral betrekking heeft op de organisatie ervan, maar hebben liever een eenvoudige representatie van wat zij zoeken.

4.3.3 Andere semantische modellen

Het Europeana Datamodel (EDM) is een veelgebruikte term in de sector echter de concrete implementatie is alleen bij Europeana te vinden en nauwelijks bij andere bronnen. EDM is vooral ontwikkeld voor de aggregatie van informatie op het Europeana platform. EDM is deels gebaseerd op CIDOC-CRM hetgeen een nog verder ontwikkeld aggregatiemodel van informatie beschrijft. De vraag die van belang is hier is of een generieke insteek verenigbaar is met een specifieke insteek, ofwel kan een dataset die volgens een algemene ontologie beschreven is worden gerelateerd aan een dataset die volgens een specifieke ontologie is beschreven, bijvoorbeeld EDM? Ja dat kan door middel van het beschrijven van een zogenaamde 'mapping'. Diverse representatie van dezelfde data kunnen worden verenigd door de semantische relatie te beschrijven tussen de verschillende representaties. Bijvoorbeeld door te definiëren dat Peuren een subklasse is van de klasse Palingvissen wordt alle informatie over Peuren uit de ene dataset gevonden via het concept Palingvissen uit de andere dataset.

RedBot kan het netwerk faciliteren met een duidelijke ontologie die is samengesteld uit standaard vocabulaire of ontologie zoals DCterms en Schema.org, en aangevuld met lokale terminologie ter beschrijving van specifieke aspecten in het domein van Fries Cultureel Erfgoed.

Bovendien kan RedBot voorzien in standaard mapping van Schema.org naar meer specifieke modellen die gebruikt worden in andere datasets.

4.3.4 Authority Data

Het gebruik van 'authority data' kan veel voordeel opleveren, vooral daar ook waar data



nog beschreven moet worden, zoals bij het digitaliseren en beschrijven van objecten. Het is daarom zinvol om een duidelijk aanbod van dergelijke data aan te bieden of ernaar te verwijzen. Hierbij kan in eerste instantie gebruik gemaakt worden van data uit o.a. DBpedia, AAT, Erfgoed-thesaurus van RCE, thesaurus van RKD of van de thesaurus van Beeld En Geluid, eventueel aangevuld met data uit eigen collectie.

'Authority data' is nodig ter vereenvoudiging van het relateren van data uit verschillende datasets. Door gebruik te maken van dezelfde 'dingen' bij de beschrijving van informatie wordt per definitie een semantische link gerealiseerd. Die 'dingen' betreffen bijvoorbeeld standaard termen of 'keywords' die binnen een breed publiek van dezelfde semantiek voorzien zijn. Er bestaat uitgebreide 'authority data' op het LOD netwerk en RedBot kan faciliteren door een adequate selectie van dergelijke data aan te bieden, verrijkt met data uit eigen collectie.

Het verdient de voorkeur dat leveranciers van datasets dezelfde authority data gebruiken ter beschrijving van hun dataset. Echter wanneer dat niet het geval is dan kan de 'authority dataset' die RedBot aanbiedt dienen als raamwerk waarmee verschillende datasets aan elkaar gerelateerd kunnen worden. RedBot 'authority data' dient als kapstok van termen voor het verenigen van data uit verschillende datasets.

4.4 Linked Open Data: Wie, Wat, Waar, Wanneer, en Hoe

Het doel van RedBot is om de Friese erfgoed collectie te verenigen en te ontsluiten als Linked Open Data. Er wordt voor Linked Open Data gekozen vanwege de voordelen die daarbij bestaan om data te verenigen en wederzijdse aanvulling te realiseren. Met name het wederzijds aanvullen van data uit verschillende bronnen zorgt voor de verrijking van het geheel, i.e. synergie. Een concreet voorbeeld is de wijze waarop wordt aangegeven wie degene is die een object heeft gecreëerd; stel een fiets die gemaakt is door Jan Janssen. In RDF termen wordt hier de volgende triple gemaakt:

```
<fiets_x> dcterms:creator <http://nl.dbpedia.org/resource/Jan\_Janssen>.
```

Dezelfde informatie in een record wordt minder gesystematiseerd weergegeven, namelijk:

```
<fiets_x> dcterms:creator "Jan Janssen"
```

Het verschil is de wijze waarop de waarde die de creator aangeeft is gesystematiseerd. In LOD termen wordt verwezen naar een publieke, autoritatieve bron zoals Dbpedia/Wikipedia. Er kan daardoor geen onduidelijkheid bestaan over Wie de creator was. In de vorm van een record wordt dezelfde waarde als string genoteerd en kan er onduidelijkheid ontstaan wanneer data uit verschillende bronnen wordt gecombineerd. Indien de string zou bestaan uit een 'identificer' naar een autoritatieve bron zou dit probleem verholpen kunnen worden!

Bovengenoemde voorbeeld illustreert de twee aspecten van standaardisatie zoals eerder beschreven;

1. Standaard modellering: Het gebruik van een gedeeld conceptueel kader. (dcterms:creator)
2. Authority data: Collectief gebruik van eenduidige concepten (nl.dbpedia).



Beide aspecten zijn van belang om invulling te geven aan de beschrijvende vectoren Wie, Wat, Waar, Wanneer, en Hoe.

Standaard modellering in RedBot sluit aan bij de wijze waarop de grote instanties in de context van RedBot data modelleren en aanbieden via hun API. Er wordt algemeen gebruik gemaakt van Dublin Core Terms t.b.v. standaardisatie van recordschema's en daarom wordt binnen RedBot daar de aansluiting gevonden; beschrijvende eigenschappen van digitale objecten worden in navolging met DCterms beschreven. Binnen DCterms zijn verschillende beschreven 'properties' relevant voor de aspecten Wie, Wat, Waar, en Wanneer. In onderstaande tabel is een (niet uitputtend) overzicht beschreven van diverse DCterms properties die relevant zijn voor de verschillende aspecten. Wat opvalt is dat het aspect 'Hoe' *niet* gerepresenteerd wordt in DCterms.



	DCterms Properties
Wie	dcterms:creator
	dcterms:contributor
	dcterms:publisher
	dcterms:audience
Wat	rdf:type
	dcterms:format
	dcterms:extent
	dcterms:subject
	dcterms:coverage
	dcterms:abstract
	dcterms:description
	dcterms:title
	dcterms:medium
Waar	dcterms:spatial
	dcterms:coverage
	dcterms:temporal
Wanneer	dcterms:date
	dcterms:created
	dcterms:issued
Hoe	

Het aspect 'Hoe' gaat in op de wijze waarop (digitale) objecten tot stand zijn gekomen; de breedte van mogelijke waarden hier is groot en vooral domeinspecifiek. DCterms is een generiek model en dus niet geschikt hiervoor; er is specifieke terminologie nodig voor het beschrijven van het creatieproces dat meestal zeer domeinspecifiek is. Een voorbeeld van een domeinspecifieke toepassing voor het 'Hoe' aspect is het



digitaliseringsproces van objecten; Het scannen van documenten bijvoorbeeld waarbij de stappen inclusief resultaten in specifieke terminologie wordt beschreven. Maar ook de creatie van een standbeeld valt onder het aspect Hoe, of de wijze waarop een kaatsbal is vervaardigd. Telkens is de wijze waarop het Hoe aspect beschreven wordt domeinspecifiek en verdient het de voorkeur om in eerste instantie het Hoe aspect over te laten aan domeinspecialisten en dus broneigenaren. In de bestaande collecties van de grote instanties wordt dit aspect in eerste instantie weinig genoemd en wordt voor informatie verstrekt over de aspecten Wie, Wat Waar en Wanneer.

RedBot kan volgen in de gekozen beschrijvingssystematiek van eigenaren van datasets en pas wanneer daartoe de wens of noodzaak bestaat kan de mapping gedefinieerd worden tussen verschillende gebruikte beschrijvingssystemen.

Standaardisatie op het gebied van authority data in RedBot betreft het gebruik van waarden voor eigenschappen die gedeeld worden via een collectieve bron. Het doel hier is in eerste instantie het realiseren van eenduidigheid in data, en in tweede instantie het verrijken van data vanwege de koppeling die gerealiseerd wordt tussen datasets. Zoals in het voorbeeld van Jan Janssen is getoond, is het gebruik van collectieve waarden voor eigenschappen belangrijk. De grote instanties in de context van RedBot zijn richtinggevend voor het gebruik van gestandaardiseerde waarden binnen RedBot. Om dit concreet te maken hieronder een indicatieve opsomming van classes die relevant zijn voor de verschillende genoemde metadata aspecten:

Wie: Personen, Organisaties, Groepen

Wat: Typering, Onderwerp, Materiaal, Omvang, Beschrijving

Waar: Locatie, Plaats

Wanneer: Tijdstip, Datum, Periode

Wat opvalt in deze opsomming is dat de aspecten Wie, Wat, en Waar allen door middel van thesauri gestandaardiseerd kunnen worden. Het aspect Wanneer wordt vooral door middel van standaardisatie van data formaat gerealiseerd en meestal niet middels thesauri. Thesauri dienen voor de identificatie en beschrijving van Personen, Plaatsen, Typering etc.

Het tijdsaspect, of het 'Wanneer' wordt gestandaardiseerd door het collectief gebruik van dezelfde notatie voor tijdseenheden. Binnen RedBot wordt geadviseerd om het XSD schema⁶ voor Datum en Tijd te hanteren.

⁶ http://www.w3schools.com/xml/schema_dtypes_date.asp



5 RedBot als service hub: gedeelde technologie

RedBot als service organisatie in het netwerk levert diensten die nodig zijn om het netwerk te ontwikkelen en in stand te houden en onderdeel te maken van de 'LOD-cloud'. Met volgende diensten kan RedBot het netwerk adviseren en ondersteunen

5.1 Standaardisatie

Een voorname dienst is het tot stand brengen van standaardisatie op het gebied van metadata modellering en data uitwisseling, waarbij zoveel mogelijk hergebruikt wordt gemaakt van ontwikkelde ontologie uit het LOD netwerk.

RedBot is geen eigenaar van de data. Maar RedBot kan wel kwaliteit van de data signaleren. Dit kan vooraf aan het toekennen van metadata bij een nieuwe collectie door als dienst een collectie-analyse te verrichten waarbij handvaten kunnen gegeven voor gestandaardiseerde invoer die voldoet aan de basisuitgangspunten (wie, wat, waar, wanneer en hoe). Achteraf (na conversie van de data naar RDF triples in RedBot kan RedBot via tools signaleren dat de data niet voldoet aan de standaard voor het 'hoe invullen van een waarde' terugkoppelen aan de collectiehouders. Dit levert schonere data op voor de collectiehouders en na opschoning betere integratie van de informatie in RedBot.

5.2 RDF mapping datasets

De conversie van data naar RDF formaat gebeurt door middel van een 'mapping' per dataset. Per dataset is een 'mapping' meestal uniek, maar wanneer mappings voor 10 of 20 datasets worden vergeleken dan bestaan er overeenkomsten waaruit generieke componenten kunnen worden herleid. Dergelijke componenten zijn herbruikbaar en vergemakkelijken het maken van een 'mapping' voor andere datasets. Op het niveau van RedBot kan gefaciliteerd worden bij het realiseren van mappings met het bijkomende voordeel dat hier optimaal gebruik gemaakt kan worden van overeenkomsten in mappings voor verschillende datasets.

5.3 Linken van datasets

Het relateren van data uit verschillende datasets heeft tot doel informatie te verenigen omdat daar toegevoegde waarde door ontstaat; Het geheel is meer dan de som der delen.

Hierboven is uitvoerig beschreven hoe informatie gerelateerd wordt door gebruik te maken van gedeelde ontologie en 'authority data'. Dergelijke relaties spelen zich vooral af op het niveau van classificatie en typering van onderwerp. Veel interessantere links tussen datasets zijn de relaties op instantieniveau, bijvoorbeeld de relatie tussen de plaats Langweer in dataset A en de plaats Langweer⁷ in de dataset van DBpedia. De link met DBpedia levert informatie op die de informatie in dataset A inhoudelijk aanvult en verbindt met het LOD netwerk.

Het vinden van links tussen datasets op instantieniveau is specialistisch werk waarbij

⁷<http://nl.dbpedia.org/resource/Langweer>



RedBot kan faciliteren. In *afbeelding 2: RDF publiceren* wordt in stap 3 onderscheid gemaakt tussen het selecteren van bronnen en data, en het uitvoeren van het 'matching' proces.

Het selecteren van relevante bronnen en data binnen die bronnen is een rol die RedBot als hub in het netwerk uitstekend kan uitvoeren vanwege het overzicht over de verschillende bronnen vanuit het perspectief van RedBot.

Het uitvoeren van het 'matching' proces is een meer complexe taak die eventueel ondersteund zou kunnen worden met een online service waarin dergelijke processen uitgevoerd kunnen worden. Hier geldt ook dat wanneer dat op een centrale plaats wordt uitgevoerd er schaalvoordelen ontstaan door hergebruik van methodiek.

Er is echter ook 'laaghangend fruit' waarmee RedBot het netwerk kan faciliteren. Wanneer data uit een dataset eenmaal geconverteerd is naar, en beschikbaar is als, RDF dan kan middels SPARQL worden gezocht naar links in andere datasets. Dergelijke SPARQL query's kunnen vanuit RedBot eenvoudig beschikbaar worden gesteld aan het netwerk.

5.4 Persistente Identifiers

5.4.1 Toelichting Persistente Identifiers

Een dataset die naar RDF wordt geconverteerd wordt voorzien van Identifiers in de vorm IRI's.

IRI staat voor Internationalized resource Identifier. Een unieke code waarmee databronnen (internationaal) uniek herkenbaar en vindbaar zijn. Op het LOD netwerk zijn IRI's de adressen van informatie die eventueel bezocht kunnen worden met een webbrowser.

Het is van groot belang dat de IRI's persistent zijn, ofwel dat zij voor lange tijd bewaard en functioneel blijven. Per slot van rekening zijn IRI's het aanknopingspunt om de link tussen data te beschrijven dus het wegvallen van (het functioneren van) een IRI betekent dat de link niet meer bestaat.

5.4.2 Rol van Redbot

RedBot kan participanten in het netwerk adviseren en faciliteren met persistente identifiers die verwijzen naar resources van participanten in het netwerk. Met name partijen die niet beschikken over een eigen domeinnaam of eigen services hebben behoefte aan een domeinnaam waarmee hun data ontsloten wordt via het web. Maar ook partijen die al RDF data hebben gepubliceerd kunnen redenen hebben om zich aan te sluiten bij een uniform systeem van identifiers binnen het domein van Fries Cultureel Erfgoed.



6 Advies

6.1 Semantiek

RedBot is de verbindende factor in Fries digitaal erfgoed. Via RedBot worden collecties van de collectiehouders verbonden met elkaar en met de informatie infrastructuur van het internet (laag 'bruikbaar' in het drielagenmodel van de Nationale Strategie Digitaal Erfgoed). Door het leggen van verbindingen vergroot je de eenduidigheid van de data uit de afzonderlijke collecties en kun je als tweede stap deze data verrijken met content uit de grote instellingen in het landschap van digitaal erfgoed Nederland en bronnen van de internet informatie infrastructuur. De meest voor de hand liggende techniek voor het leggen van verbindingen is de techniek Linked Data, via de standaard RDF.

Het gebruik van de techniek Linked Data betekent niet dat de collectiehouders zelf de data als (RDF) triples dienen aan te leveren. In '150122 Adviesrapport Inpassing Friese Erfgoedhub in Frysk Kultuerhus' adviseert de werkgroep de strategie van RedBot, nl. de data 'zo hoog mogelijk' binnen de nationale infrastructuur te harvesten. Zo hoog mogelijk wil zeggen: collectiehouders brengen de eigen data onder in een van de vijf domeinen en via een van de vijf grote domeinen (Beeld&Geluid, KB/ Delpher, Nationaal Archief, RCE en DANS/ KNAW) haalt RedBot de data weer binnen. RedBot converteert de data na harvesting van de data uit een van de vijf grote instellingen naar (RDF) triples. Collectiehouders metadateren zelf, bij voorkeur uitgaande van Datamodellering van RedBot. Serviceorganisatie RedBot biedt diensten aan dit proces voor de collectiehouders zo goed mogelijk te laten verlopen.

Advies semantiek:

- Gebruik RedBot als verbinder van de Friese (erfgoed) collecties conform het drielagenmodel van de Nationale Strategie Digitaal erfgoed;
- Gebruik Linked Data techniek via standaard RDF;
- Registreer de gedigitaliseerde materialen, waar mogelijk, op landelijk of provinciaal niveau;
- Laat het converteren van de data naar (RDF) triples binnen RedBot te verzorgen.

6.2 Datamodel

Efficiëntie uit het verbinden van collecties van diverse collectiehouders via Linked Data wordt gehaald door het gebruik van standaarden door collectiehouders en goede afspraken tussen collectiehouders en RedBot. RedBot wil als knooppunt diverse soorten data (denk aan boekbeschrijvingen, beschrijvingen van museale objecten of beschrijvingen uit beeldbanken) integraal ontsluiten. Om deze diverse data met elkaar te verbinden en betekenis te kunnen geven, adviseert RedBot een datamodel dat uitgaat van vijf basisuitgangspunten: wie, wat, waar, wanneer en hoe.

De eerste vier (wie, wat, waar, wanneer) zijn collectie-inhoudelijk, de vijfde (hoe) betreft de wijze waarop (digitale) objecten tot stand zijn gekomen. RedBot realiseert deze modellering via standaardisering en het hergebruik van gestandaardiseerde, breed gebruikte metadata modellen of ontologie. Deze efficiëntie wordt het meest gerealiseerd door per project bij digitalisering en invoeren van beschrijvende metadata (beschrijvingen)



alvast voor te sorteren op gebruik van de standaarden en de authority files zoals DBpedia. Als collectiehouders bij de invoer al rekening kunnen houden met invoer volgens standaarden en volgens links naar authority files, dan verloopt de onderlinge uitwisselbaarheid en uitlijning met bv. een DBpedia veel soepeler hetgeen resulteert in nauwkeurige integrale collectieontsluiting voor de eindgebruiker.

Advies datamodel:

- De werkgroep adviseert om het toekennen van metadata te systematiseren en onderling afspraken hierover te maken tussen RedBot en de collectiehouders.
- De voorkeur gaat uit naar metadatering volgens de 'Dublin Core Terms' standaarden en het datamodel vooraf met de collectiehouder af te stemmen (zie pa. 4.4) Voor dit vooraf of tijdig afstemmen van metadatering biedt RedBot diensten aan als onderdeel van RedBot Service Organisatie (zie pa. 6.3).

6.3 RedBot Serviceorganisatie

Fries digitaal erfgoed in een informatie infrastructuur via een verbindingslaag (RedBot) is meer dan het implementeren en toepassen van techniek. Een Serviceorganisatie kan collectiehouders ondersteuning bieden in het integraal ontsluiten van Fries digitaal erfgoed via het knooppunt RedBot. RedBot kan een waardevolle rol spelen in deze ondersteuning, zowel voor de collectiehouders voor goede integratie van de collecties naar het Friese knooppunt, als ook voor kwaliteit van de eigen metadata.

Advies RedBot serviceorganisatie:

- De werkgroep adviseert RedBot de rol op zich te laten nemen als Service Organisatie. RedBot beschikt over kennis op het gebied van digitaal erfgoed en heeft de beschikking over een groot netwerk voor digitaal erfgoed. Technische diensten die RedBot kan leveren, verhogen de kwaliteit van de collecties van de collectiehouders (en daarmee van Fries digitaal erfgoed) (Standaardisering, pa. 5.1) en maken dit digitale erfgoed duurzaam toegankelijk (Persistent identifiers pa. 5.4). Het helpt de collectiehouders de data volgens de standaarden en de Datamodellering onder te brengen in de vijf grote domeinen.
- De werkgroep adviseert vorm en inhoud van de RedBot serviceorganisatie te laten beschrijven en ter goedkeuring voor te leggen aan de stuurgroep RedBot.



7 Samenstelling werkgroep RedBot

Leden werkgroep:

- Luc de Vries, Tresoar (materiedeskundige);
- Hans Laagland, Tresoar (materiedeskundige);
- Olav Kwakman, Tresoar (materiedeskundige);
- Roland Cornelissen, Metamatter (extern deskundige);
- Bertus Douwes, Bibliotheekservice Fryslân (projectleider).

