Nederlandse Parels van Linked Data Toepassingen

Gems of Dutch Linked Data Applications

Platform Linked Data Nederland

PlatformLinkedData.nl

Voorwoord

Is Linked Data meer dan een buzzwoord? Het Platform Linked Data Nederland nam de proef op de som door een oproep te plaatsen voor de beste Linked Data-toepassing. Vrijwel gelijktijdig werd ook een Europese prijs ingesteld voor de beste Linked Data-toepassing. Wij konden de handen ineen slaan en zo dongen de Nederlandse inzendingen voor het Platform Linked Data Nederland ook direct mee naar de Europese prijs.

Het resultaat mag er zijn! Zestien inzendingen met toepassingen van de farmaceutische industrie tot de culturele sector en van basisregistraties tot vreemdelingenbeleid. En van aansprekende, eenvoudige toepassingen met een ongekende winst tot aan zeer complexe innovatieprojecten die zich nog moeten bewijzen.

De Nederlandse jury, bestaande uit Jos van Hillegersberg (Universiteit Twente), Rob van de Velde (Geonovum, initiatiefnemer van het Platform Linked Data Nederland), Rinke Hoekstra (VU Amsterdam), Erwin Folmer en Linda van den Brink (trekkers van het Platform Linked Data Nederland) heeft de volgende prijswinnaars geselecteerd.

In de categorie 'Linked Open Data' heeft de LOD Laundromat van de VU Amsterdam de eerste prijs gekregen. In de categorie 'Linked Enterprise Data' kwam de NXP Enterprise Datahub van NXP Semiconductors & SEMAKU als beste uit de bus.

Ook in de European Linked Data Contest sleepte de Nederlandse inzending van NXP de eerste prijs binnen in de categorie 'Linked Enterprise Data'. In de categorie 'Open Data' ging de award naar OpenPHACTS. OpenPHACTS is een Europees project in de farmaceutische industrie met een stevige Nederlandse inbreng. Ook in Europees verband dus een fantastisch resultaat voor de Nederlandse projecten.

In deze bundel zijn de Nederlandse inzendingen bij elkaar gebracht. De diversiteit van sectoren waarin Linked Data haar weg vindt, heeft ons blij verrast. Wij nemen u daar graag in mee.

Platform Linked Data Nederland

Erwin Folmer
Geonovum, Kadaster, Universiteit Twente

Linda van den Brink
Geonovum

Platform Linked
Data Nederland

De LOD Laundromat

VU Amsterdam

Linked Open Data Application Award 2015, The Netherlands

European ELDC Award 2015

Linked Open Data Toepassing Award 2015, Nederland

Europese ELDC Award 2015

The biggest collection Linked Open Data in the world

The LOD Laundromat provides access to all Linked Open Data (LOD) in the world. It does this by crawling the LOD cloud, and converting all its contents in a standards-compliant way, removing all data stains such as syntax errors, duplicates, and blank nodes.

Using and finding Linked Data takes time and effort. Not all available Linked Data is clean, standard and easy to use. Many datasets contain syntax errors, duplicates, or are difficult to find. We offer one single download location for Linked Data, and publish the Linked Data in a consistent simple (sorted) N-Triple format, making it easy to use and compare datasets.

De grootste verzameling Linked Open data in de wereld

De LOD Laundromat biedt toegang tot alle Linked Open Data in de wereld. De Laundromat doorzoekt het web op Linked Data en wast alle vlekken, zoals syntaxfouten, dubbelingen of lege velden eruit. Wat overblijft zijn schone, conform de Linked Data-standaarden opgestelde datasets die je zo van de waslijn kunt plukken.

Het vinden en gebruiken van Linked Data kost veel tijd en vraagt de nodige inspanning. Niet alle beschikbare Linked Data zijn schoon, gestandaardiseerd en makkelijk te gebruiken. Veel datasets bevatten syntaxfouten, duplicaten, of zijn moeilijk te vinden. De LOD Laundromat lost dit op door één downloadlocatie te bieden voor Linked Data, de data op te schonen en gestandaardiseerd beschikbaar te stellen. Dat gebeurt in een consistente eenvoudige (gesorteerde) N-Triple-indeling, waardoor je de datasets gemakkelijker kan gebruiken en vergelijken.

http://bit.ly/1rc9vxl

De LOD Laundromat

Hoe het werkt...

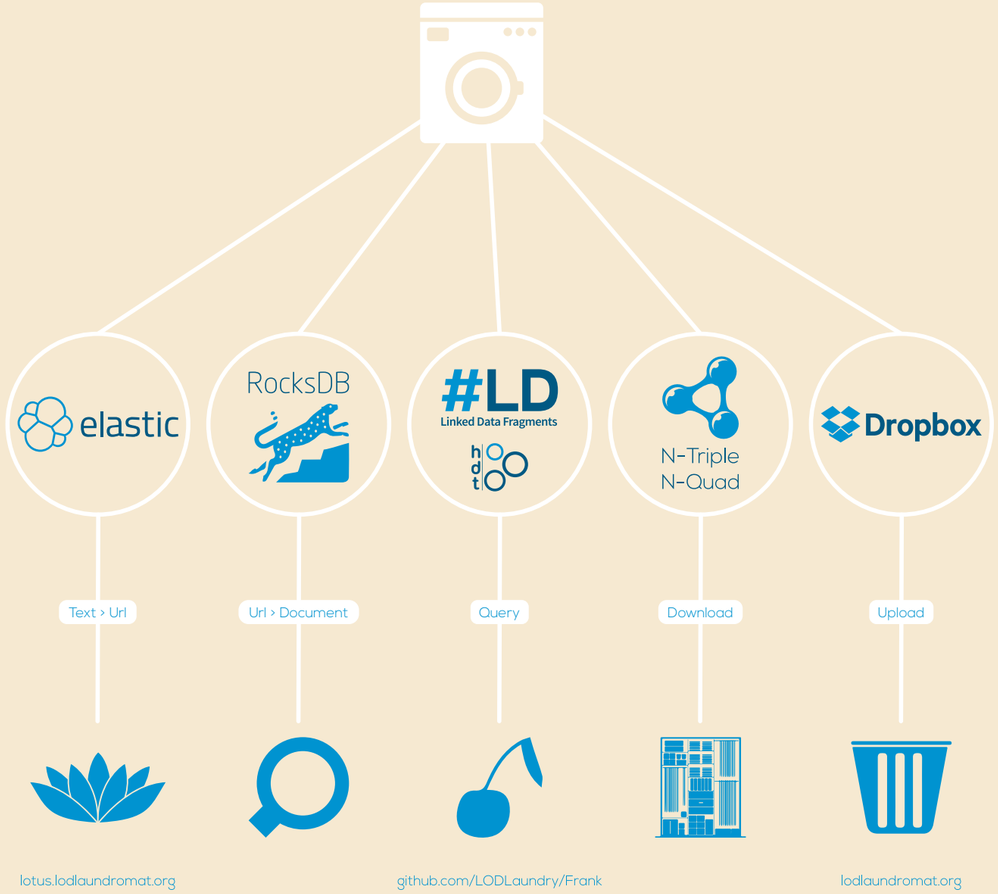
De LOD Laundromat is de eerste Linked Open Data-API die uniform én grootschalig is. Om dit te realiseren maakt de LOD Laundromat-architectuur gebruik van de laatste technologische ontwikkelingen op het gebied van datastreaming, dataparsing, dataopslag en querying. Door gebruik te maken van innovatieve dataopslagmethodes, zoals HDT en LDF, is het mogelijk om data op ongekende schaal online te bevragen.

De LOD Laundromat neemt niet alleen werk uit handen bij het zoeken en vinden van Linked Data, maar helpt ontwikkelaars ook om hun data zelf beter beschikbaar te stellen. Zo worden de niet-standaards conforme eigenschappen van iedere dataset uitvoerig beschreven in metadata en is het mogelijk om schoongewassen tussenversies van de data voor eigen gebruik te downloaden.

De Laundromat publiceert de gegevens als gezippede, gesorteerde N-Triples en N-Quads, of als geïndexeerde en gecomprimeerde HDT-bestanden. Het biedt ook toegang tot Triple Pattern Fragment-API's. De herkomst en de VoiD-meta-data zijn toegankelijk via het SPARQL-endpoint van de Laundromat.

De LOD Laundromat-aanpak is domeinonafhankelijk en wordt ook buiten de computerwetenschappen ingezet, bijvoorbeeld binnen de lexicografie.

Zie ook http://lodlaundromat.org



LinkedLegislation

KOOP

Linked Open Data Application Award 2015, The Netherlands

Linked Open Data Toepassing Award 2015, Nederland

Unleashing the power of open data

In 2012 the Publication Office of the Netherlands launched a programme to publish laws and regulations as Linked Data. Using the standard Juriconnect, data is linked between different collections of public information, such as case law, policy, and notices. The Dutch laws can be downloaded as Linked Open Data.

The Government produces and publishes laws and regulations and makes these available as open data. Publishing this information as Linked Open Data immediately shows the relationship between data in various authentic sources. Thus, it diminishes the time lost on figuring out these relationships. It also makes it much easier to assess the impact of a change in the law.

In the programme, much attention has been given to dealing with validity of laws and regulations. As a result, even thirty years from now, you will be directed exactly to that part of the law that was valid at the requested time. To be able to find older documents that have no URI, a 'link extractor' has been developed. The link extractor automatically detects links to laws and regulations, case law, parliamentary documents and EU directives.

Ontketent de kracht van open data

In 2012 is het Kennis- en Exploitatiecentrum Officiële Overheidspublicaties (KOOP) gestart met een programma om wet- en regelgeving als Linked Data te publiceren. Met behulp van de standaard Juriconnect, wordt Linked Data gemaakt uit verschillende collecties van overheidsinformatie, zoals jurisprudentie, beleid en kennisgevingen. De Nederlandse wetten zijn al compleet als Linked Open Data te downloaden.

De overheid produceert en publiceert wet- en regelgeving en stelt deze beschikbaar als open data. Het publiceren als Linked Open Data van deze informatie maakt dat de relaties tussen data in verschillende authentieke bronnen direct in beeld zijn. Hierdoor gaat minder tijd verloren met het uitzoeken van deze relaties. Ook is het nu veel eenvoudiger om de impact van een wetswijziging in kaart te brengen.

In het programma is veel aandacht besteed aan het omgaan met geldigheid van wet- en regelgeving. Als gevolg hiervan leidt de Linked Data je - ook over dertig jaar nog - precies naar dat deel van de wetgeving dat op het gevraagde tijdstip van kracht was. Om ook oudere documenten zonder URI's te kunnen vinden is een 'link extractor' ontwikkeld. Deze detecteert automatisch links naar wet- en regelgeving, jurisprudentie, parlementaire stukken en EU-richtlijnen.

http://linkeddata.overheid.nl

LinkedLegislation

Hoe het werkt...

In het 'portal' linkeddata.overheid.nl zijn vier functies beschikbaar:

1. Linktool voor wet- en regelgeving waarmee gestandaardiseerde links naar wet- en regelgeving uit wetten.nl kunnen worden gemaakt volgens de Juriconnect-standaard.
2. Brongegevens waarbij per bron de ingelezen objecten en aantal relaties opvraagbaar zijn, inclusief een doorklik naar de spiegelpagina.
3. Spiegelpagina waarin de gerelateerde links bij wet- en regelgeving of andere overheidsinformatie wordt getoond
4. LAB-omgeving waarin experimentele functies worden getoond, zoals:
   - de linkextractor (zie ook http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2626521)
   - de linktool-jurisprudentie
   - de officiële linktool-publicaties
   - de call-backfunctie voor aanroep van de linktool in websites en toepassingen

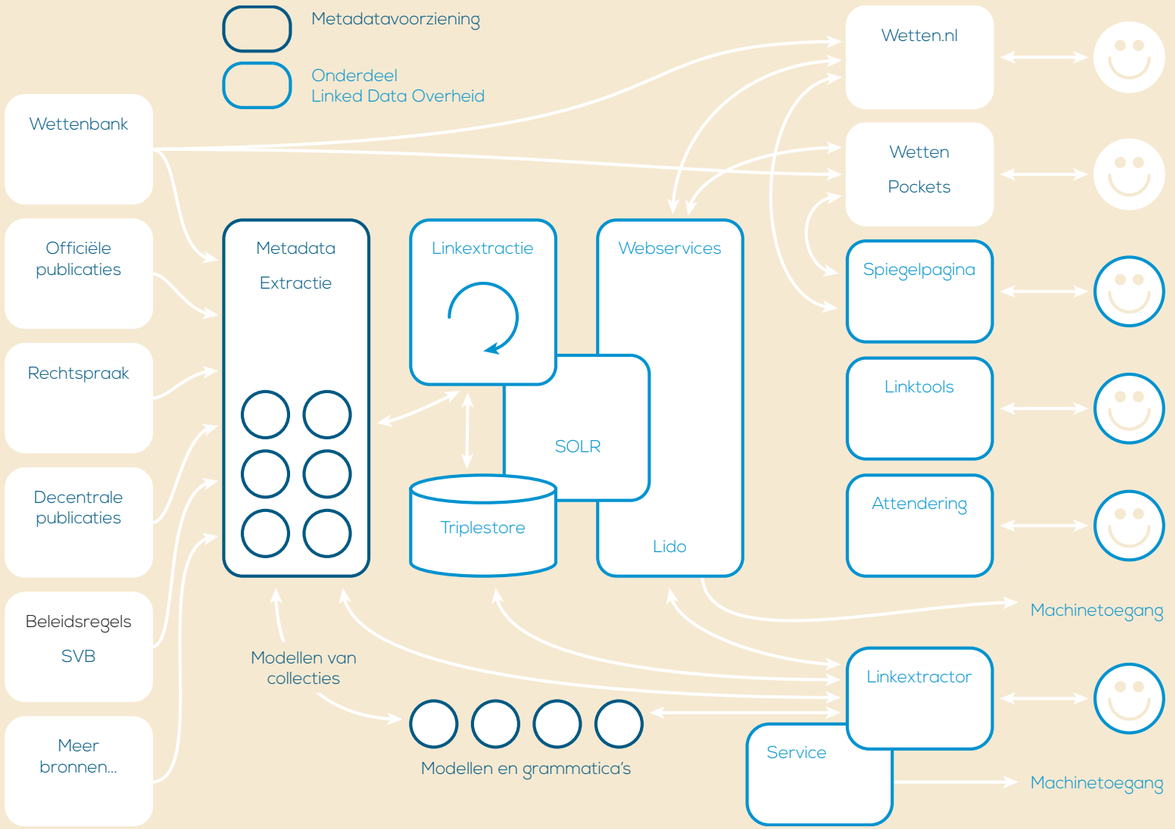
Widgets en webservices

De linktool voor wet- en regelgeving en de spiegelpagina zijn ook als 'widget' beschikbaar. Een widget is een webpagina die makkelijk is te integreren in een eigen toepassing. Men kan de widgetversie van de linktool en spiegelpagina oproepen door in de URL het woord 'front' te vervangen door 'embedded', bijvoorbeeld http://linkeddata.overheid.nl/embedded/portal/linktool-bwb. Een specifieke vorm van de linktoolwidget is de 'call-backwidget' die bijvoorbeeld vanuit een macro in Microsoft Office (Word, Excel) kan worden aangeroepen om links naar wet- en regelgeving te maken.

De Linked Data-voorziening wordt volledig als webservice aangeboden. Overheidsorganisaties kunnen eigen toepassingen aan deze webservices koppelen voor het ophalen van gerelateerde links of het maken van links. Om ook voor de interne informatievoorziening van Linked Data gebruik te maken kunnen overheden een eigen instantie van de Linked Data-omgeving toepassen: 'Linked Data in-a-box'.

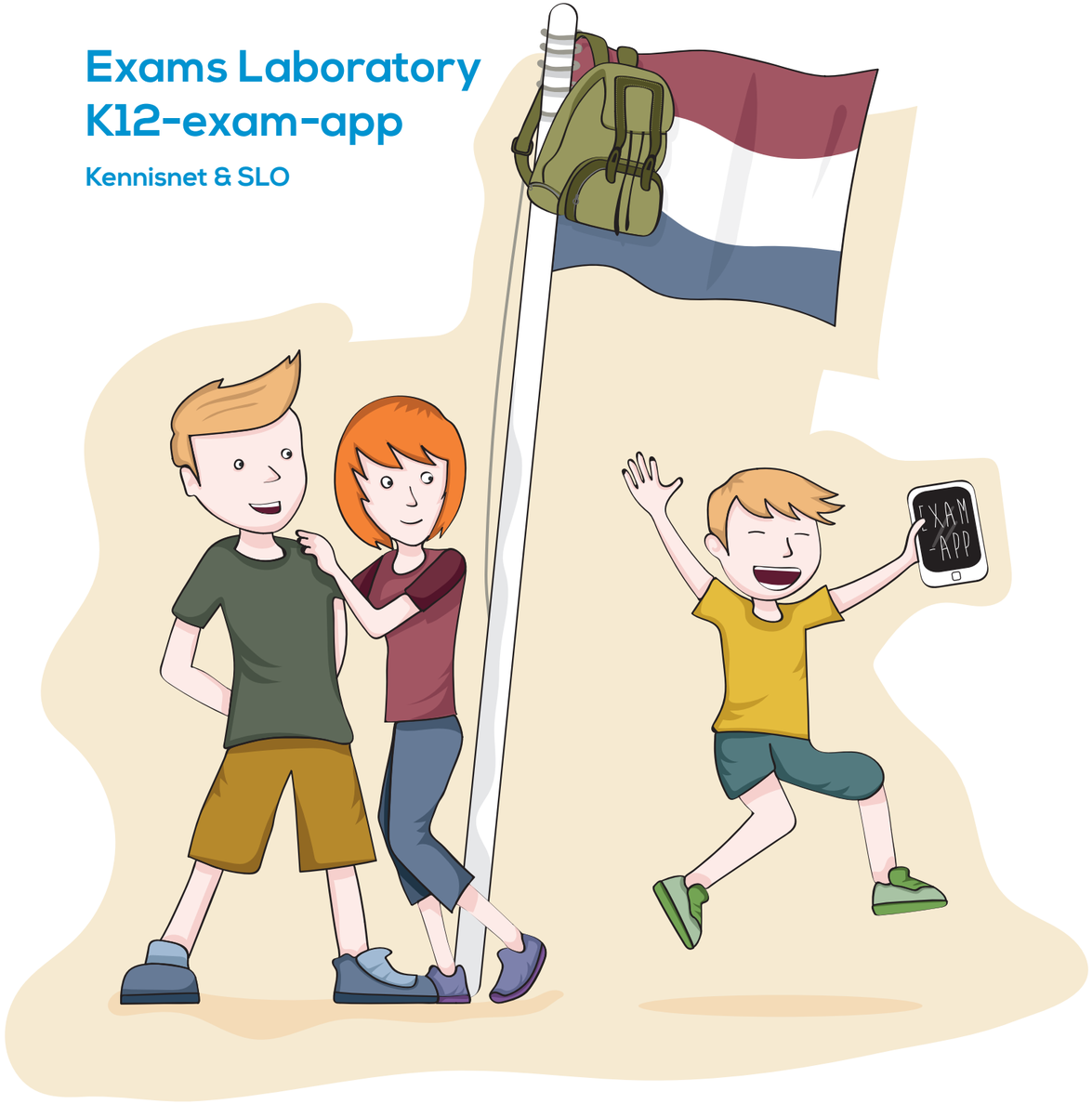
Standaarden

1. De Linked Data-voorziening ondersteunt het maken van links naar wetten.nl volgens de verwijsstandaarden Juriconnect 1.0 en 1.3.
2. De Linked Data-voorziening ondersteunt het opvragen van informatie conform de standaarden Juriconnect 1.0 en 1.3. Daarnaast kunnen voor het opvragen ook de 'eigen' identifiers van de ingelezen bronnen worden gebruikt, zoals een BWB-ID, ECLI of OP-nummer.
3. De Linked Data-voorziening maakt voor alle objecten een http-URI, de LIDO-identifier.
4. De Linked Data-voorziening voorziet in een 'RDF about' voor elk opgenomen informatieobject. De about-pagina kan worden opgevraagd met de http-uri van het betreffende informatieobject. De http-uri van een object kan worden verkregen via de linktool of uit de spiegelpagina.



Exams Laboratory
K12-exam-app

Kennisnet & SLO



Linked Open Data Application Award 2015, The Netherlands

Linked Open Data Toepassing Award 2015, Nederland

Tailored training aimed at the final exams

Being able to train on specific parts of the final exams for secondary education. Tailoring educational material on the learning curve and knowledge of an individual pupil. Linked data supports achieving these aspirations.

In a laboratory setting, questions from numerous CITO-graduation assignments have been brought together, including national exam scores (psychometric data). Each question is associated with the chapters and paragraphs of more than thirty teaching methods. By entering the scores of the practice exams the pupil gets insight into his or her progress and can immediately see which material he should re-study. Teachers can follow progress per pupil and the class as a whole.

This project has led to the publication of various Linked Open Datasets in education, including the conceptual framework for education. Access to this conceptual framework is facilitated by a practical API (see www.kennisnet.nl/diensten/onderwijsbegrippenkader/gebruik-obk). All educational publishers have cooperated, by publishing the detailed structure of the textbooks. As this is competitive information, the application works with a mixture of closed and open linked metadata. The privacy aspect also weighed heavily as actual scores from pupils are used. The application uses the actual scores from pupils. A special security audit has been done to ensure the security of personal data.

Studeren op maat voor je eindexamens

Gericht kunnen trainen voor je havo-eindexamen. Onderwijsstof toesnijden op de leercurve en kennis van een individuele leerling. Linked data helpt deze ambities verwezenlijken.

In de proeftuineindexamens zijn diverse CITO-eindexamenopgaves verzameld, inclusief de landelijke examenscores (psychometrische data). Elke opgave is gekoppeld aan de hoofdstukken en paragrafen van meer dan dertig lesmethodes. Door zijn eigen scores van de oefenexamens in te voeren krijgt de leerling inzicht in zijn voortgang én wordt direct duidelijk welke leerstof hij nog eens door moet nemen. De docent kan per leerling of voor de gehele klas de voortgang volgen.

Dit project heeft geleid tot de publicatie van diverse linked datasets in het onderwijs, waaronder het onderwijsbegrippenkader. Op dit onderwijsbegrippenkader is ook een praktische API ontwikkeld (zie www.kennisnet.nl/diensten/onderwijsbegrippenkader/gebruik-obk). Alle educatieve uitgeverijen hebben meegewerkt door hun methodestructuur beschikbaar te stellen. Dit is concurrentiegevoelige informatie. De applicatie werkt daarom met een mix van linked open en gesloten metadata. Ook het privacy-aspect woog zwaar, omdat er met echte leerlingdata werd gewerkt. Om de veiligheid van persoonlijke gegevens te waarborgen is een security audit uitgevoerd.

http://bit.ly/1rHtJzM

Exams Laboratory K12-exam-app

Hoe het werkt...

Het Exams Laboratory bestaat uit een datalaag en een applicatielaag met een API als verbindende schakel. De datalaag bestaat uit twee componenten: een tool om de Linked Open Data te managen (RNA-omgeving van de firma InfoProjects) en een triplestore met SPARQL-endpoint (Virtuoso). De applicatielaag is gebouwd met PHP en bestaat uit een data-cache voor de Linked Open Data, een database voor gebruikersgegevens, een functionele/logische laag en een userinterface. De applicatie is voor authenticatie en autorisatie gekoppeld met de Kennisnet Federatie (SAML). De applicatie is tevens gekoppeld met twee eindexamentrainingsproviders. Op de 'heenweg' worden gebruikers doorgelaten via SSO (Kennisnet Federatie) en op de 'terugweg' wordt een score meegegeven die in de database van de applicatie wordt opgeslagen.

De Linked Open Data omvat in RDF:

- Kernprogramma's van acht havo-vakken van SLO (uit het OnderwijsBegrippenKader)
- Gegevens van meer dan vijftig eindexamens (opgaven, koppeling aan kernprogramma, psychometrische gegevens van Cito)
- Inhoudsopgaven van meer dan dertig lesmethoden.

De database van de applicatie omvat:

- Gebruikersgegevens van meer dan vierduizend leerlingen op bijna honderd scholen
- Bijna vijftigduizend ingevoerde scores (gemaakte eindexamenopgaven).

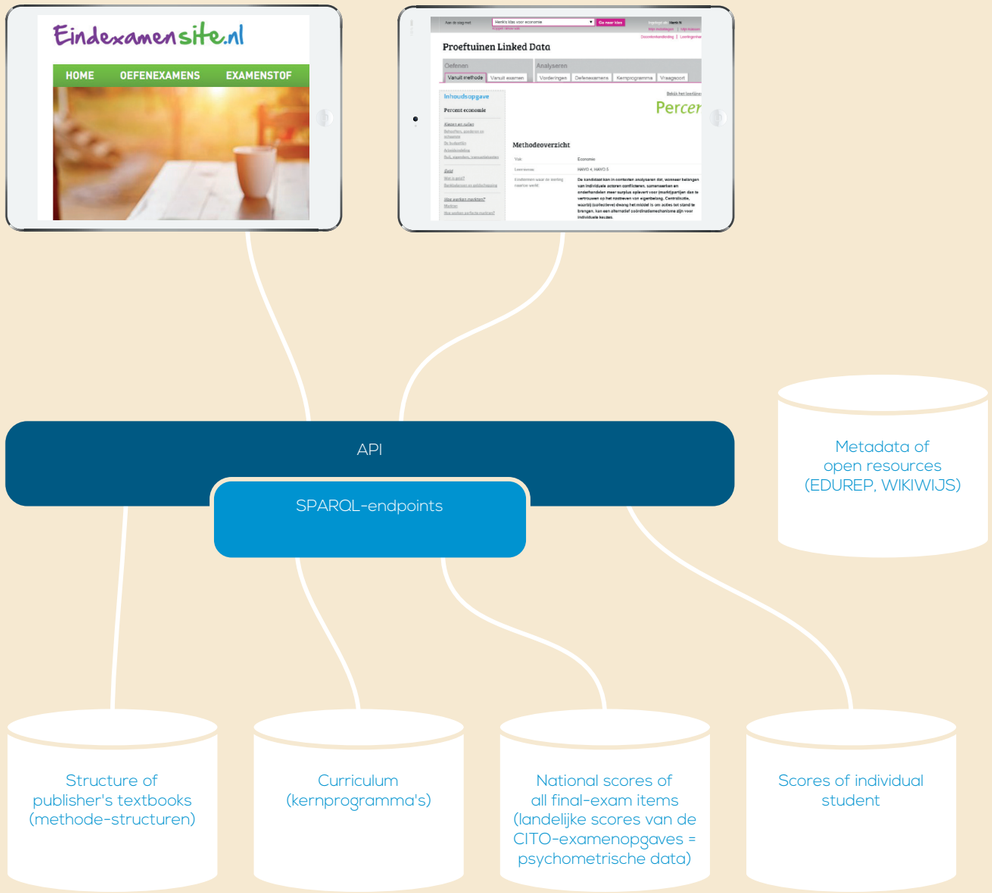
De database bevat privacygevoelige informatie en is niet voor derden toegankelijk.

Uitdaging

De belangrijkste technische uitdaging is de performance van de toepassing. De RDF-data en de SPARQL-query's kunnen zo complex worden dat sommige query's tien seconden duurden – een voor gebruikers onacceptabele responstijd. Dit is opgelost door de query's te optimaliseren en door een data-cache in de applicatie in te bouwen. Eén keer per dag werd deze cache geheel vernieuwd door het aflopen van alle SPARQL-query's. Maar zelfs mét caching bleken de query's die de applicatie moest doen, nog zo complex dat de responsetijd van sommige analyseoverzichten enige seconden bedroeg.

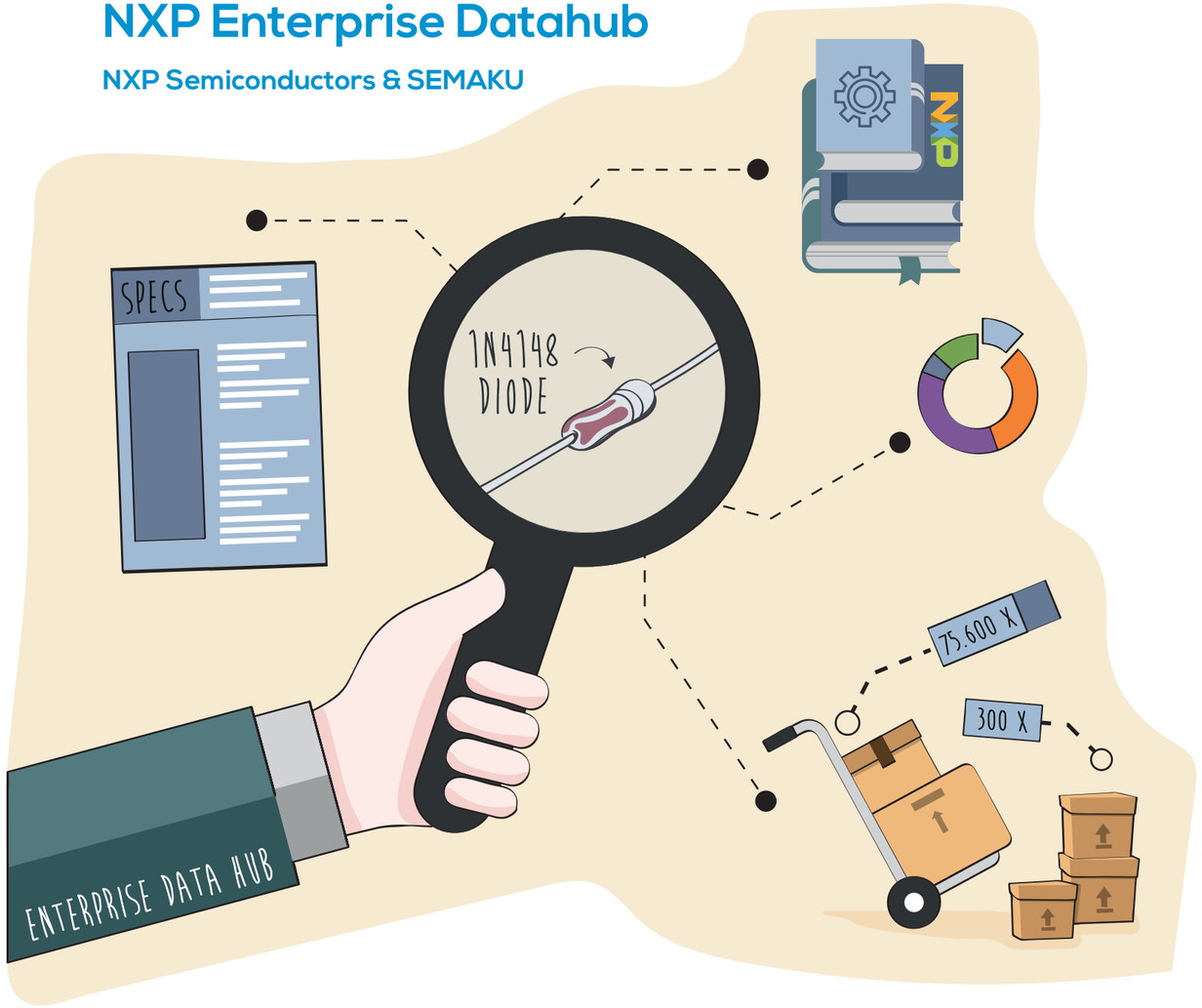
Daarnaast is het curriculum-model redelijk complex: de kernprogramma's van SLO zijn modelmatig gezien vier niveaus diep en kennen verschillende contexten. Dat stelt zwaardere eisen aan de modellering. Een begrip als 'water' bestaat bijvoorbeeld zowel in de context van scheikunde, als in die van biologie en natuurkunde. Dat leidt tot complexere query's in de API-laag.

https://proeftuinexamens.kennisnet.nl/homepage/index | http://bit.ly/1YgBVBl



NXP Enterprise Datahub

NXP Semiconductors & SEMAKU



Linked Enterprise Data Application Award 2015, The Netherlands

European ELDC Award 2015 for Linked Enterprise Data

Linked Enterprise Data Toepassing Award 2015, Nederland

Europese ELDC Award 2015 voor Linked Enterprise Data

Your product data always up to date

NXP has a portfolio of over 20.000 products across a wide range of functions and technologies. Information about each of these products is stored and managed in various internal systems. How to keep this information up to date in different systems? NXP tackles this challenge with Linked Data.
Information on products needs to be communicated both internally and to the customer via multiple channels and formats: web, mobile and print. Data is thereby scattered and duplicated across numerous applications and databases. The outcome is that people are unable to find what they are looking for, or find conflicting information.
With the Enterprise Data Hub, data from these different databases can be integrated using the Linked Data principles and standard web technologies. For example, with the help of shared identifiers, data about the electrical characteristics can be easily merged with data from the product lifecycle. All available data about a product can be viewed in a single knowledge graph. And as the identifiers are hyperlinks, users can access the data simply by opening the links in their web browser. NXP has thus implemented Linked Data as a vital backbone for their information services.

Productdata altijd en overal bijgewerkt

NXP biedt meer dan 20.000 producten aan met een breed scala aan functies en technologieën. Informatie over elk van deze producten wordt in diverse interne systemen opgeslagen en beheerd. Hoe zorg je er dan voor dat informatie in alle verschillende systemen wordt bijgewerkt? NXP lost het op met Linked Data.
Informatie over producten wordt op verschillende manieren en via meerdere kanalen intern en naar de klant gecommuniceerd: bijvoorbeeld via internet, mobiel en print. Hierdoor raken data verspreid en wordt informatie gedupliceerd in meerdere applicaties en databases. Dit leidt ertoe dat mensen niet kunnen vinden wat ze zoeken of tegenstrijdige informatie vinden.
De Enterprise Data Hub, een data-integratietechnologie, zorgt ervoor dat data uit verschillende databases geïntegreerd kunnen worden via Linked Data-principes en standaard web-technologieën. Dankzij 'shared identifiers' kan men bijvoorbeeld data over de elektrische eigenschappen van een product samenvoegen met gegevens over de levenscyclus van het product. Alle beschikbare gegevens over dit product kunnen daardoor worden bekeken in één overzicht. Doordat de identifiers hyperlinks zijn, kunnen gebruikers de data met één muisklik bekijken in hun webbrowser. Voor NXP zijn Linked Data daarmee een niet meer weg te denken ruggengraat voor hun informatiediensten.

http://bit.ly/21pq3io

NXP Enterprise Datahub

Hoe het werkt...

De NXP Enterprise Datahub haalt data op uit bronsystemen, transformeert deze naar RDF en laadt ze vervolgens in de RDF Graph Database, van waaruit ze conform Linked Data-principes worden gepubliceerd. Deze benadering heeft de voorkeur boven het gebruik van een R2RML-wrapper om een virtuele SPARQL-endpoint mee te maken, omdat veel van de bronsystemen een dergelijke toegang niet toestaan of niet relationeel zijn.

De brondata bestaan uit diverse XML- en CSV-formaten en kennen verschillende updatefrequenties. In het eenvoudigste geval kunnen data automatisch worden opgehaald. In andere gevallen worden berichten verspreid via de Enterprise Service Bus (ESB) en implementeren we de transformatie in het kanaal. Van XML-bronnen transformeren we de data naar RDF (om precies te zijn: RDF/XML of Trix), gebruikmakend van XSLT of XQuery. Voor CSV-bronnen brengen we de CSV naar een Java-object, vervolgens naar XML, gevolgd door XSLT.
De data worden in de database geladen met het SPARQL 1.1 Graph Store HTTP Protocol.

In een aantal gevallen gebruiken we de RDF Graph Database ook als eerste opslagbron voor de data. In deze gevallen zet een applicatie de data direct in de store met behulp van SPARQL 1.1 Query Language en SPARQL 1.1 Update via ofwel SPARQL 1.1 Protocol of gebruikmakend van opgeslagen SPARQL-procedures.

Beveiliging

NXP moet ook rekening houden met de informatiebeveiliging. Om te voorkomen dat interne data per ongeluk publiek toegankelijk worden, zijn strikte filtermechanismen opgezet. Deze filteren de openbare subsets uit interne data. Dit gebeurt met behulp van SPARQL 1.1 Federated Query samen met SPARQL 1.1 Update waar de 'business rules' worden gevangen in SPARQL-language.

Waar mogelijk heeft NXP gebruik gemaakt van bestaande standaard vocabulaires als Dublin Core, SKOS, FOAF en Schema.org. Omdat er een aantal voor de sector benodigde termen en concepten ontbraken, is ook een NXP-vocabulaire ontwikkeld met mappings (zoals rdfs:subClassOf) naar externe vocabulaires. Op deze manier zijn er voldoende mogelijkheden om de data te beschrijven en kunnen externe gebruikers de betekenis van begrippen gemakkelijker nagaan.

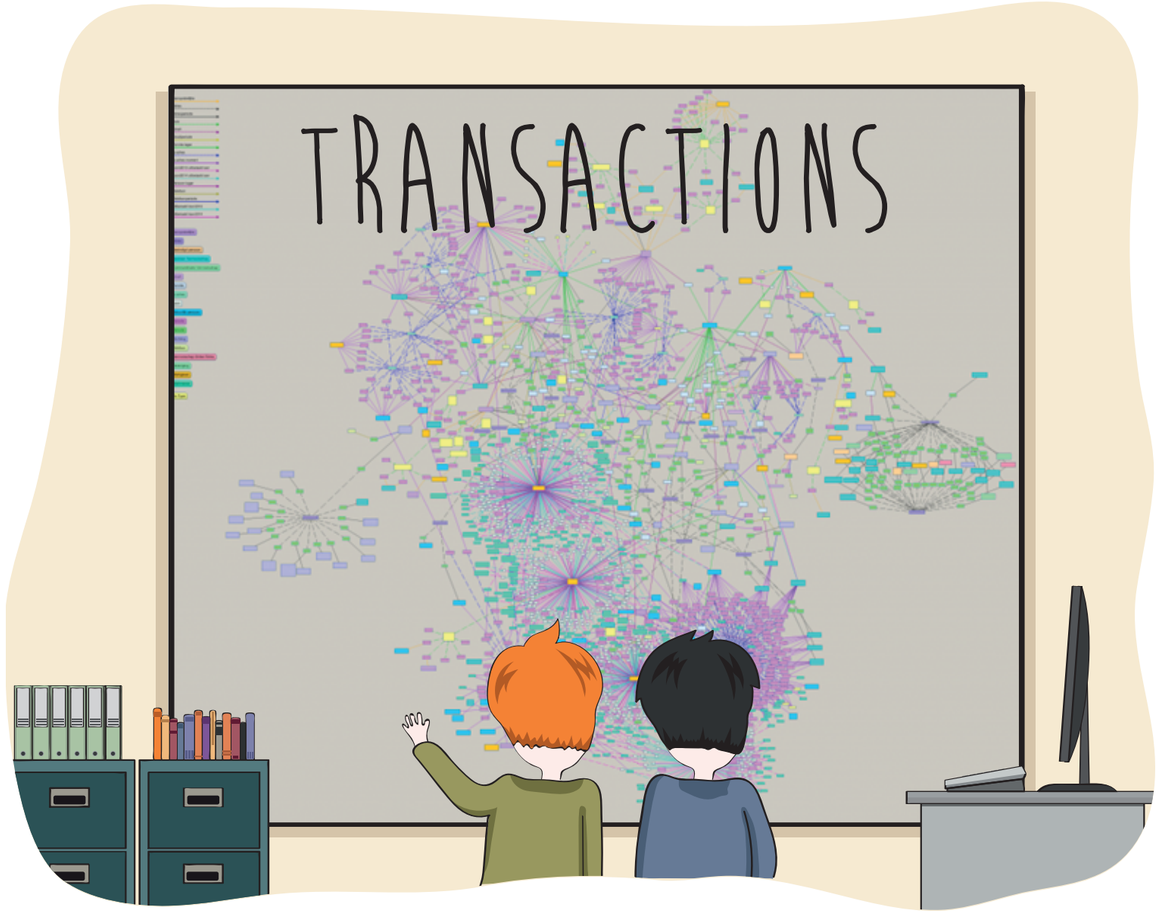
De openbare data zijn toegankelijk op http://data.nxp.com. Door dereferencing van de 'entity URI's biedt NXP ook een SPARQL endpoint en dataextracten. Om te kunnen volgen wie welke informatie gebruikt, is een API-sleutel vereist.

Uitdaging

De grootste uitdaging in dit project was het behoud van consistentie van data bij het laden via een ESB. Een RPC-benadering kan tot inconsistentie leiden als er iets fout gaat. Men wilde elke databasetransactie laten corresponderen met één bericht. De technische uitdaging hier zat hem in hoe je dit werkend maakt via HTTP waarbij er geen adequaat PATCH-format bestaat om met RDF-quadsdata om te gaan. Gelukkig konden we de graafbeperkingen in de store opheffen en een extensie van het SPARQL 1.1. Graph Store http Protocol definiëren voor de PATCH-operatie die alleen de graaf in het bewuste bericht wist en de andere grafen intact laat.

Fraud detection

Belastingdienst



Linked Enterprise Data application 2015
The Netherlands

Linked Enterprise Data Toepassing 2015
Nederland

The power of association

Linked data gives an unexpected boost to the detection of fraud and fraudsters. Detective work that now costs an inspector two weeks of time, can be done in five minutes when using Linked Data.

Linked data uses identifiers. An identifier is a unique feature of a thing or a person. This unique feature is used to refer to the thing or the person. However, these identifiers can also be used as nodes of data. Different data sources can thus be combined. So the key register Addresses and Buildings can be combined with the Trade register as the same identifier for the address is used. The BSN-number (or the RSIN for companies), is also a powerful node.

Fraud experts at the tax office use a visualisation of the Linked Data nodes. Any node can be clicked, after which you can access the nodes that link to that particular item. This gives the fraud expert infinite possibilities to carry out its analysis. The visualisation of the network supports the thinking of the fraud expert. He retains overview but can also discover where there may be a new lead.

De kracht van associatie

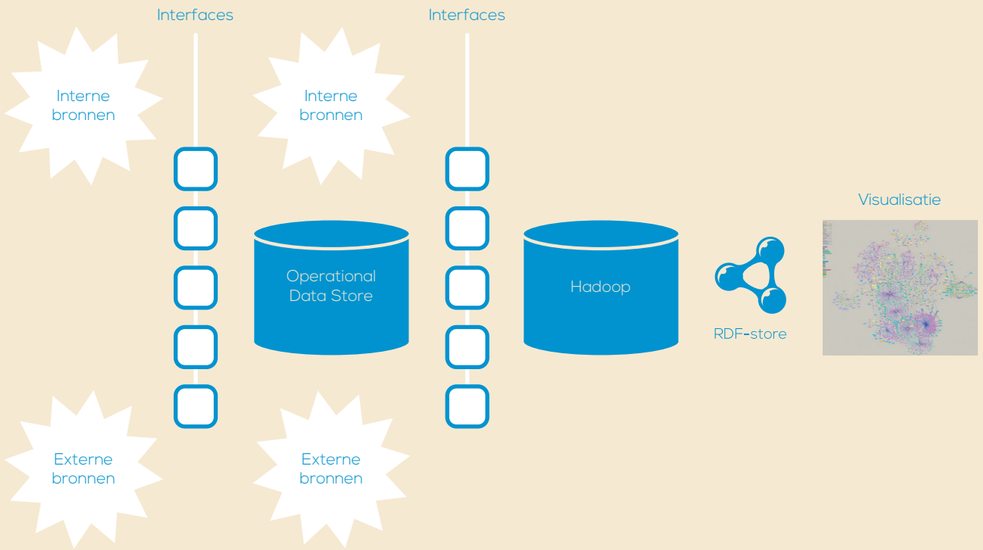
Linked data geeft een onverwachte boost aan het opsporen van fraude en fraudeurs. Speurwerk dat een inspecteur nu twee weken kost, kan met behulp van Linked Data worden gedaan in vijf minuten.

Linked data maakt gebruik van 'identifiers'. Een identifier is een uniek kenmerk van een ding of een persoon. Dit unieke kenmerk gebruikt men om het ding of de persoon mee aan te duiden. Deze identifiers kan men echter ook gebruiken als knooppunten van data. Verschillende databronnen kunnen op die manier met elkaar worden gecombineerd. De Basisregistratie Adressen en Gebouwen kan men bijvoorbeeld combineren met het Handelsregister, omdat dezelfde identifier voor het adres wordt gebruikt. Ook het BSN (of het RSIN voor bedrijven) is een krachtig knooppunt.

De fraude-experts bij de Belastingdienst gebruiken in hun werk een visualisatie van de gelinkte dataknooppunten. Ieder knooppunt kan worden aangeklikt, waarna je toegang krijgt tot de daaraan gelinkte knooppunten. Dit geeft de fraude-expert oneindig veel mogelijkheden voor het uitvoeren van zijn analyse. De visualisatie van het netwerk ondersteunt het denken van de fraude-expert. Hij behoudt overzicht maar kan tevens ontdekken waar er mogelijk een nieuw verband is.

Fraud detection

Hoe het werkt...

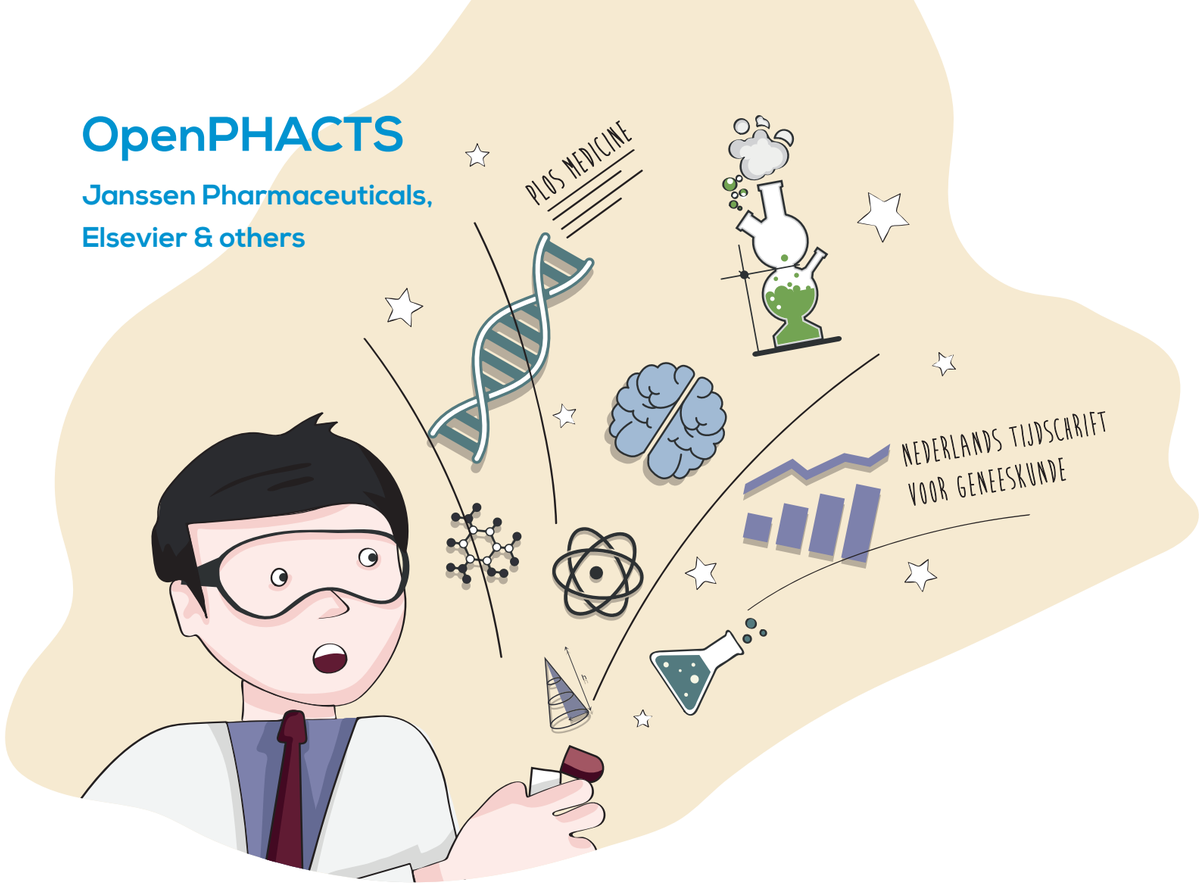


Ter ondersteuning van het speurwerk naar fraude wordt gebruik gemaakt van de interface AllegroGraph. Deze interface wordt alleen gebruikt om te lezen. Input en update van data vindt plaats op de RDF-store. Om de relaties tussen gegevens te leggen wordt gebruik gemaakt van standaarden voor identifiers.

De belangrijkste uitdaging in dit project lagen op het organisatorische vlak: toestemming om te investeren in deze technologie en mensen om het werk te doen. Inhoudelijke uitdaging ligt op het vlak van betekenis geven: welke knooppunten en welke relaties zijn relevant voor medewerkers die toezicht uitvoeren?

OpenPHACTS

Janssen Pharmaceuticals, Elsevier & others



Speeding up pharmaceutical research

For pharmaceutical research many and diverse sources of information are available. Simultaneously searching and analysing data from diverse biomedical categories, was hardly possible or extremely time consuming. With the help of semantic web standards and Linked Data, this has changed.

Versnellen van farmaceutisch onderzoek

Voor farmaceutisch onderzoek zijn veel en diverse informatiebronnen beschikbaar. Het simultaan doorzoeken en analyseren van data in diverse biomedische categorieën was niet mogelijk of extreem tijdrovend. Dankzij semantische webstandaarden en Linked Data, is dat nu anders.

European ELDC Award 2015
for Linked Open Data

Linked Enterprise Data application 2015
The Netherlands

Europese ELDC Award 2015
voor Linked Open Data

Linked Enterprise Data Toepassing 2015
Nederland

OpenPHACTS provides a semantic platform that uses Linked Data and semantic web standards to facilitate pharmaceutical research. Thanks to the attention that this public private partnership has given to data ontologies and vocabularies, data licensing and copyright, it is now possible to search a large amount of public biomedical information in one workflow.

The scientific value of OpenPHACTS is significant. It has already led to new analytical methods, such as graph analyses on diverse biomedical data (see for instance www.euretos.com). Computer analysis on the available data will lead to new insights and significantly speed up research on diseases and pharmaceuticals. OpenPHACTS is also of economic value. The project has integrated public biomedical data and in the process, structured it, by giving much attention to uniform vocabularies and ontologies. Moreover: the Linked Open Data system can easily be extended with proprietary data.

OpenPHACTS biedt een semantisch platform waarin Linked Data en semantische webstandaarden worden toegepast om farmaceutisch onderzoek te ondersteunen. Dankzij de aandacht die deze publiek-private samenwerking heeft besteed aan ontologieën en woordenlijsten, gegevenslicenties en auteursrecht, is het nu mogelijk om een grote hoeveelheid biomedische open data in een workflow te doorzoeken.

De wetenschappelijke waarde van OpenPHACTS is wezenlijk. Het heeft al tot nieuwe analysemethoden geleid, zoals graph analyses op diverse biomedische databronnen (zie www.euretos.com). Computeranalyses van beschikbare gegevens zullen leiden tot nieuwe inzichten en onderzoek naar ziekten en geneesmiddelen aanzienlijk versnellen. OpenPHACTS is ook van economische waarde. Het project heeft ervoor gezorgd dat open biomedische data zijn geïntegreerd en al doende gestructureerd, door veel aandacht te geven aan uniforme woordenlijsten en ontologieën. Bovendien is het Linked Open Data-systeem eenvoudig uit te breiden met gesloten datasets.

http://www.openphacts.org

OpenPHACTS

Hoe het werkt...

OpenPHACTS is een semantisch platform dat uit een aantal componenten bestaat: User Interface & Applications, Core API, Linked Data Cache, Identity Resolution Service, Identity Mapping Service, Domain Specific Services en Databronnen.

Op basis van de onderzoeksvragen en beschikbaarheid zijn de volgende databases opgenomen in Open PHACTS: ChEBI, ChEMBL, ConceptWiki, DisGeNET, DrugBAnk, ENZYM, FAERS, Gene Ontology, neXTProt, Uniprot, Wikipathways. Dit resulteert in een RDF-gegevensset van meer dan 3 miljard triples.

Bij de realisatie van OpenPHACTS moesten enkele cruciale aspecten worden uitgewerkt: gegevensontologieën en woordenlijsten, licenties en auteursrecht. De naamgeving van stoffen in de biomedische wereld is rommelig en dubbelzinnig. De chemische wereld biedt in veel opzichten zelfs nog meer uitdagingen door dubbelzinnigheid en degeneratie in haar identificatiesystemen. Deze complexiteit van de namespace is opgelost met een systeem waarin individuele concepten die de biologische concepten omvatten (genen, eiwitten, drugs/chemische stoffen, ziekten,...) evenals de sociale concepten (auteurs, artikelen, datasets,...) op het niveau van het individuele concept worden gehouden. Liever dan alle gegevensverstrekkers wereldwijd te dwingen om op een standaard manier te verwijzen naar deze afzonderlijke concepten, gebruikt OpenPHACTS on-the-fly-identiteittoewijzing om verschillende termen en IRI's voor dezelfde fysieke entiteit te combineren.

Het voordeel van deze aanpak is dat verschillende regels kunnen worden toegepast bij een query. OpenPHACTS stelt niet één specifieke woordenschat verplicht voor een bepaald semantisch type (gen, eiwit, drug,...), maar geeft aanbevelingen over hoe gegevens het best worden getoond binnen het systeem.

OpenPHACTS beveelt ook het gebruik van openbare vocabulaires en identifierschema's aan, zoals NCBO's BioPortal (http://bioportal.bioontology.org/ontologies), EBI's Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup/) en BridgeDB (http://bridgedb.org/). De complexiteit van het verwerken van chemische samengestelde gegevens wordt aangepakt met behulp van het al bestaande ChemSpider-platform.

Een praktische kwestie bij de integratie van meerdere gegevensbronnen is auteursrecht. Het internationaal recht is niet uniform met betrekking tot auteursrecht van gegevens. Dit kan leiden tot tal van praktische problemen. Een expliciete verklaring van auteursrecht en een licentie zijn cruciaal als je data wilt kunnen delen en hergebruiken. Als resultaat van de op dit vlak gepleegde inspanningen heeft OpenPHACTS nu kristalhelder auteursrecht en licenties voor alle opgenomen gegevensbronnen.

User Interface & Applications

Core-API Het eerste prototype van het Open PHACTS-platform had alleen een SPARQL-eindpoint waar geïntegreerde gegevens konden worden opgevraagd. Om daar op aan te sluiten moest elke 'drug discovery'-applicatie zelf de vereiste SPARQL-query's schrijven en goede kennis hebben van de beschikbaar gestelde gegevens. Om aansluiting eenvoudiger te maken is de Core-API geïntroduceerd in de architectuur. De Core-API biedt een reeks methoden die toepassingen standaard kunnen oproepen. Je hoeft daarvoor geen eigen SPARQL-query's meer te schrijven.

Linked Data Cache waarbij is gekozen voor centrale opslag van de data in een LDC omwille van betrouwbaarheid en prestaties.

Domain specific services Er is al een verscheidenheid aan farmacologische services beschikbaar in specifieke domeinen die betrouwbaar zijn en goed werkende implementaties kennen. Een goed voorbeeld is het toewijzen van verbindingen op basis van chemische structuren in plaats van namen in ChemSpider. In plaats van deze toepassing opnieuw te maken, gebruikt OpenPHACTS deze en andere bewezen services.

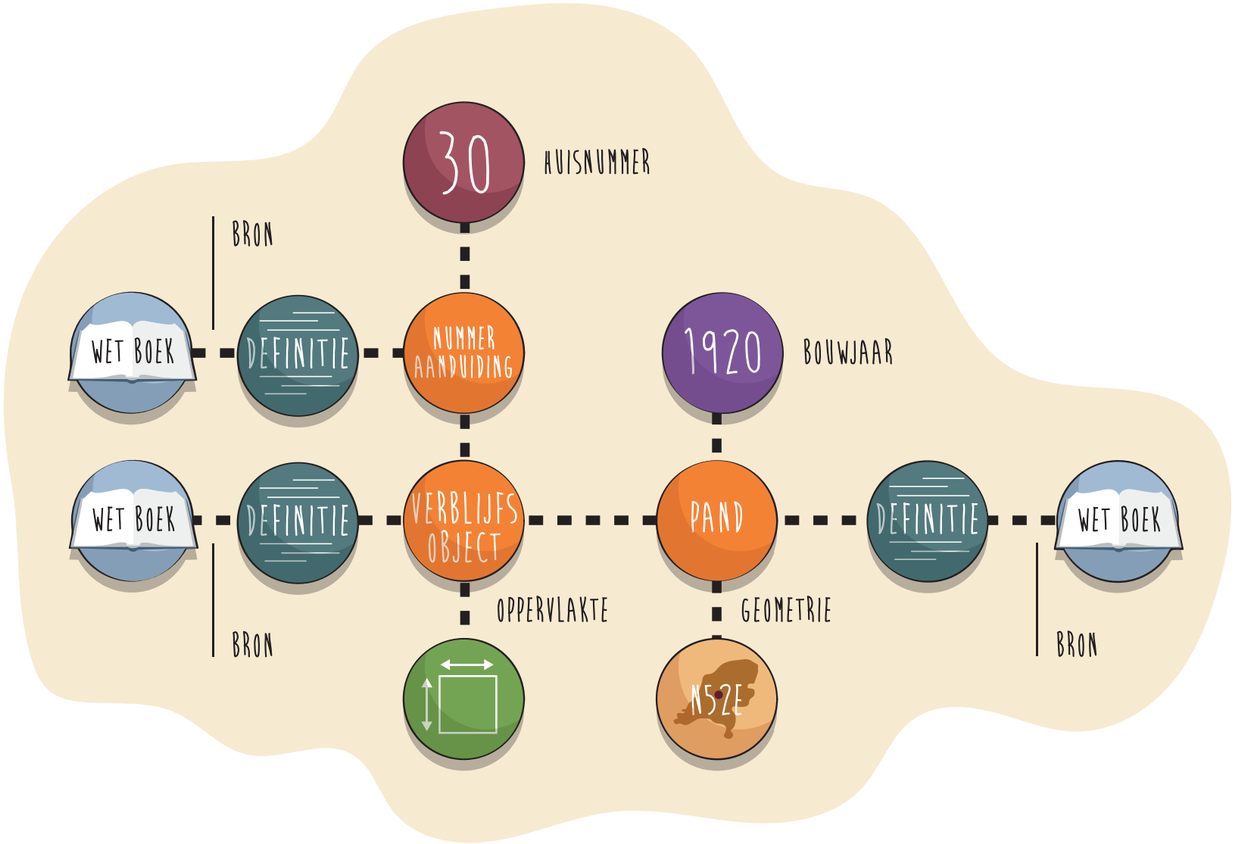
Databronnen, waarbij OpenPHACTS met bestaande oorspronkelijke RDF-gegevensbronnen werkt die worden gehost in een OpenLink Virtuoso triplestore; dit stimuleert partijen om RDF-gegevens te blijven verstrekken.

Identity Resolution Service die van door gebruikers ingevoerde beschrijvingen van entities (vrije tekst) 'bekende entities' maakt, met een gedefinieerde URI. Deze bekende entities kunnen vervolgens worden gebruikt in zoekopdrachten.

Identity Mapping Service. Gelijkwaardigheid is contextafhankelijk. Wanneer je op zoek bent naar targets waarmee een bepaalde chemische stof reageert, zullen sommige gegevensbronnen mappings hebben gemaakt naar gen-identifiers in plaats van naar eiwit-identifiers: in dergelijke gevallen kan het aanvaardbaar zijn dat gebruikers de gen- en eiwit-identifiers als vergelijkbaar beschouwen. Echter in andere situaties kan dit juist niet aanvaardbaar zijn. Het platform moet deze dynamiek met betrekking tot gelijkwaardigheid ondersteunen. Dat gebeurt door de identifierlinks niet hard te coderen in de datasets, maar de links door te leiden naar de IMS. Tijdens de query worden dan de juiste links bepaald.

Linked Data Theatre

Kadaster & Ordina



Key registers as Linked Data

The Linked Data Theatre makes Key Registers available using the LOD standards and using HTML5 features for a human-friendly presentation. Spin-offs of the original service have been used in banking, tax and infrastructure.

Key registers are central sources of data Dutch governmental organisations have to use. The Key Registers Adresses and Buildings and Topography are examples. Much information in the Key Registers is public. Via the Linked Data Theatre, this data is made available as Linked Open Data.

For the description of the concepts in the registers a knowledge system is built in accordance to the web-standard SKOS. In addition, a link has been laid from the data to the laws and regulations they relate to, thus defining their meaning. For describing the data, web standards such as SHACL are used. For describing models RDFS and for describing classes OWL/DL. For the description of metadata DCAT is used. This creates a linked 'insert' on the data, describing exactly the meaning of concepts, the coherence between data and showing how the data are collected and how reliable, timely, and complete they are.

Basisregistraties als Linked Data

Het Linked Data Theater maakt Basisregistraties toegankelijk met behulp van LOD-standaarden en gebruikt HTML5 om de data te presenteren. Spin-offs van de service zijn te vinden in de bancaire sector, belastingen en infrastructuur.

De basisregistraties zijn centrale databronnen waarvan Nederlandse overheidsorganisaties verplicht gebruik maken. Voorbeelden zijn de basisregistratie 'Adressen en Gebouwen' en 'topografie'. Veel informatie uit de basisregistraties is openbaar. Via het Linked Data Theater zijn deze data als Linked Open Data gepubliceerd.

Voor de beschrijving van de begrippen in de registratie is een kennissysteem gebouwd conform de webstandaard SKOS. Bovendien is een relatie gelegd tussen de data en de wet- en regelgeving waaruit de betekenis van deze data te herleiden is. Voor het beschrijven van de data worden webstandaarden zoals SHACL gebruikt, voor het beschrijven van modellen RDFS, en voor het beschrijven van klassen OWL/DL. Voor de beschrijving van metadata is DCAT gebruikt. Zo ontstaat een gelinkte 'bijsluiter' over de data die precies beschrijft wat haar betekenis is, die de samenhang tussen gegevens inzichtelijk maakt en die laat zien hoe de data zijn ingewonnen en hoe betrouwbaar, tijdig en volledig zij zijn.

Data: http://brk.kadaster.nl | http://bag.kadaster.nl | http://tax.kadaster.nl | http://brt.kadaster.nl

Linked Data Theatre

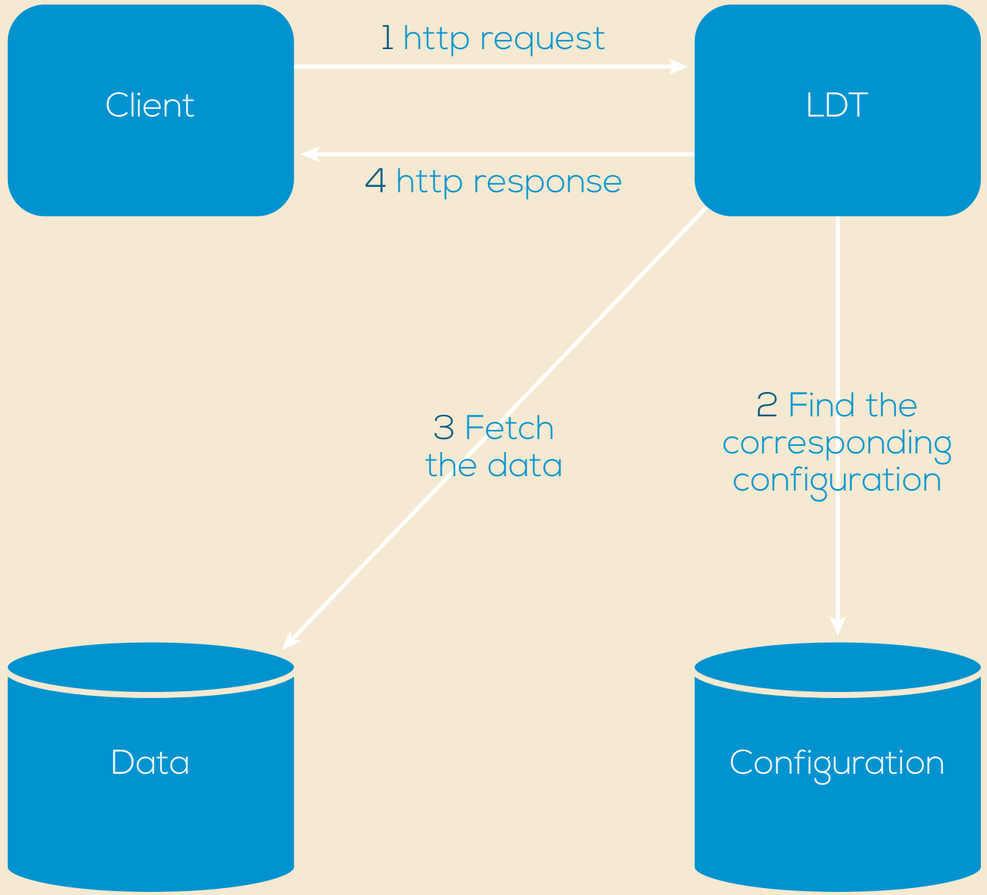
Hoe het werkt...

Het Linked Data Theater is een opensourceapplicatie die het mogelijk maakt om Linked Data zowel beschikbaar te stellen voor machines en app-bouwers (in Turtle, JSON-LD, RDF/XML), als deze op een mensvriendelijke manier te tonen in HTML. Het Linked Data Theater onderscheidt zich met een toepassing waarmee men eenvoudig definities van de onderliggende begrippen opstelt. De definities zelf worden ook in Linked Data opgesteld. Het Theater combineert een aantal opensourceraamwerken om de informatie te tonen, te weten: Twitter-bootstrap (voor een responsive userinterface), D3.js (voor het tonen van grafische weergave van Linked Data) en leaflet.js (voor het tonen van geografische gegevens).

Op basis van een http-request zoekt het Linked Data Theater naar de corresponderende configuratie in een Linked Data-configuratietriplestore. Als er geen configuratie beschikbaar is, wordt een 'default' configuratie gebruikt. De configuratie omvat een of meer SPARQL-query's die worden gebruikt om data van (een of meer) triplestores te halen. De resultaten worden teruggegeven aan de client, in het door de client aangegeven formaat (tabel, kaart, grafisch).

Linked Data Theatre Features

- Responsive UI based on Bootstrap
- Multi-format: HTML, XML, JSON, spreadsheet, text, PDF, SVG en graphml support
- Supports secure access via https and authentication using standard authentication protocols.
- Zero-coding: the configuration consists completely of RDF triples
- Supports Spatial and Graphical representations
- Open source available on Github, GNU license
- Multi-channel: PC, tablet and smartphone support
- Capable of accessing multiple SPARQL endpoints



CultuurLINK

Spinque

Linking with a strategy

CultuurLINK is a public service for cultural heritage organisations that helps them integrate their collections. While several tools exist to perform the alignment of datasets, they are not always easy to apply. As a result, people end up creating ad hoc solutions on a case-by-case basis. Each use case again requires (programming) knowledge. Moreover, as the linking process is not defined, it is impossible to replicate it.

CultuurLINK supports the iterative process of linking. It comprises the analyses of datasources, defining mapping rules and evaluating results. This results in a linking strategy that not only generates the links, but also provides the provenance and can be reused to connect updated or new datasets.

The Spinque LINK technology can be applied in various industries and for various problems. CultuurLINK has been developed to support the cultural heritage community with the alignment of their vocabularies. It is available at http://cultuurlink.beeldengeluid.nl. With CultuurLINK, collection administrators themselves link their terminology resources to large thesauri that are authoritative for the community. Based on their current sources, they can then gain access to external data that enrich their own collections by, for example, additional background information or multilingual descriptions.

Linken met een strategie

CultuurLINK is een publieke dienst waarmee instellingen voor culturele erfgoed zelf hun collecties kunnen integreren. Hoewel er verschillende programma's bestaan voor het linken van datasets, zijn deze niet altijd gemakkelijk toe te passen. Het gevolg is dat er keer op keer ad-hoc oplossingen worden ontwikkeld. Hierdoor vraagt elke 'use case' opnieuw (programmeer)kennis. En doordat het ad-hoc linkproces niet eenduidig wordt vastgelegd, is het niet reproduceerbaar.

CultuurLINK ondersteunt het iteratieve proces van het leggen van links. Het omvat het analyseren van databronnen, het definiëren van mappingregels en het evalueren van de resultaten. Het resultaat is een linkstrategie die niet alleen de links genereert, maar ook de herkomst aangeeft en kan worden hergebruikt om bijgewerkte of nieuwe datasets te verbinden.

De Spinque LINK-technologie kan in diverse sectoren en voor verschillende problemen worden ingezet. CultuurLINK is ontwikkeld om de erfgoedsector te ondersteunen met het afstemmen van haar vocabulaires. Het is beschikbaar op http://cultuurlink.beeldengeluid.nl. Met CultuurLINK kunnen collectiebeheerders zelf eigen woordenlijsten mappen op grote thesauri die leidend zijn in de erfgoedsector. Op basis van hun eigen bronnen krijgen ze zo toegang tot externe gegevens die hun eigen collecties verrijken. Bijvoorbeeld met meer achtergrondinformatie of meertalige beschrijvingen.

CultuurLINK has already been used by the NIOD Institute for War, Holocaust and Genocide Studies to link their terminology with subject terms to a large audiovisual thesaurus (GTAA) from the Netherlands Institute for Sound and Vision. The resulting links connect, among other things, a collection of Dutch historic newsreels with photographic collections of the NIOD. An application developed on top of this data can now suggest photographs to the newsreels.

CultuurLINK is al gebruikt door het NIOD, instituut voor oorlogs-, holocaust- en genocidestudies, om hun terminologie te mappen op de grote audiovisuele thesaurus (GTAA) van het Netherlands Instituut voor Beeld en Geluid. Dit heeft onder andere geleid tot een link tussen historische nieuwsuitzendingen met de beeldcollectie van het NIOD. Een applicatie op basis van deze data doet beeldsuggesties bij nieuwsitems.

http://www.comsode.eu/index.php/2015/07/linked-open-images | http://cultuurlink.beeldengeluid.nl

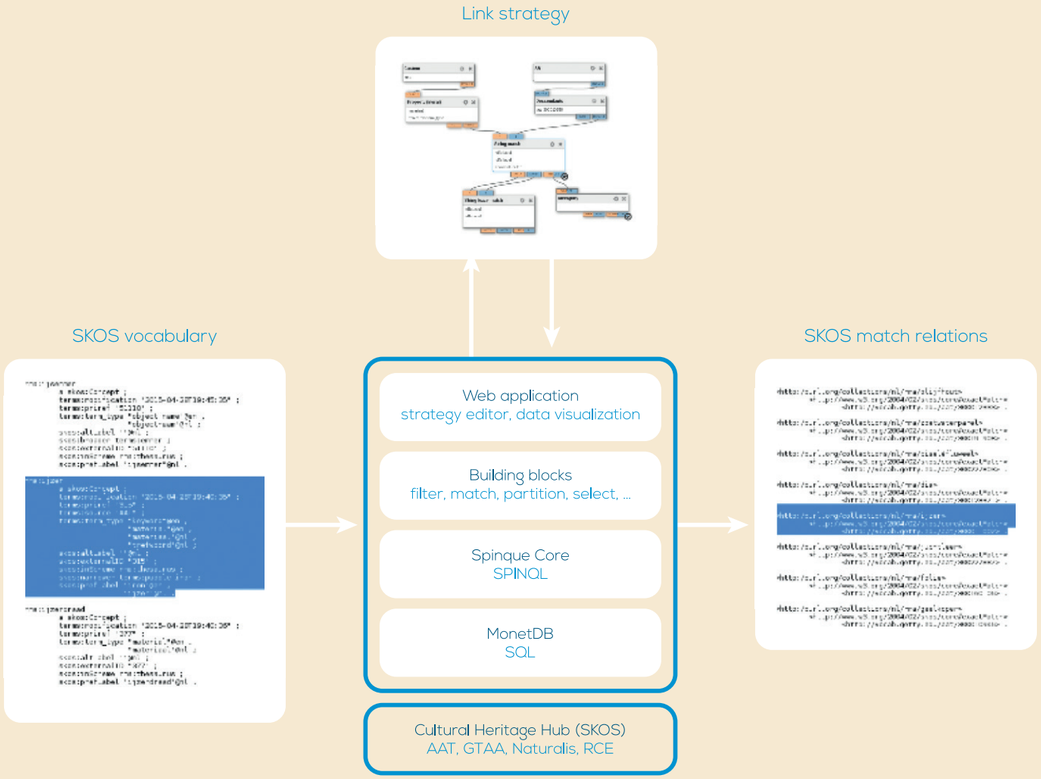
CultuurLINK

Hoe het werkt...

CultuurLINK helpt de gebruiker woordenlijsten te stroomlijnen die zijn gemodelleerd in SKOS. De dienst neemt de beschikbaar gestelde SKOS-woordenlijst en mapt deze op één van de grote thesauri in de hub van de service. De uitvoer is de linkstrategie, samen met de bijbehorende set van links, gepresenteerd als 'SKOS match relations'. De linkstrategie is een verklarende beschrijving van het proces van de mapping en de herkomst (provenance), en kan altijd opnieuw worden uitgevoerd, bijvoorbeeld wanneer onderliggende woordenlijsten verder zijn ontwikkeld.

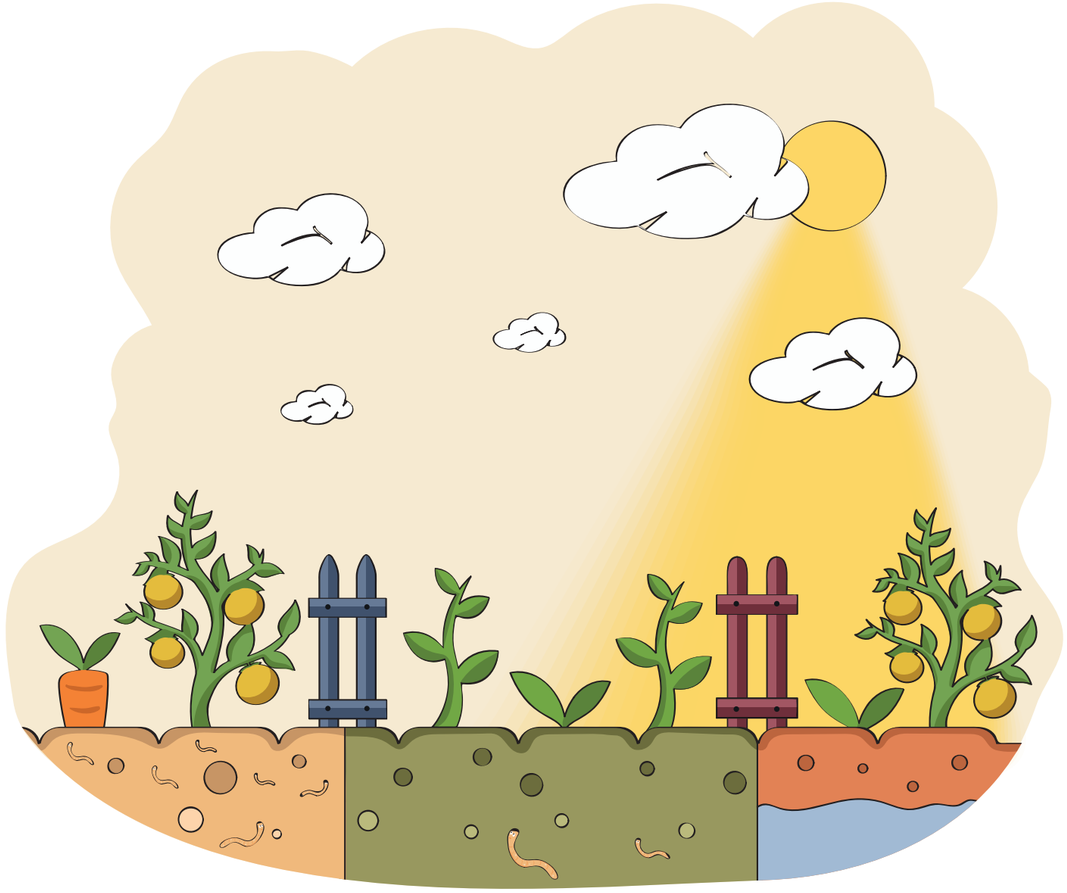
CultuurLINK is gebouwd op basis van Spinque's zoektechnologie die zoeken over meerdere gegevensbronnen mogelijk maakt. In dit geval wordt gezocht naar koppelingen tussen meerdere gegevensbronnen. CultuurLINK biedt bouwstenen om de teksten in RDF-bronnen te vergelijken op basis van exacte en fuzzy overeenkomsten. Ook kan de structuur in de RDF-bronnen worden meegenomen om de resultaten te verbeteren. In de interface bepaalt de gebruiker stap voor stap welke teksten en structuur worden gebruikt om de links te vinden. Omdat de uitvoer van elke stap wordt gevisualiseerd, kan de gebruiker onmiddellijk beoordelen of de gekozen linkstrategie leidt tot bevredigende resultaten, of dat hij moet worden verfijnd.

In de backend worden de functionele eisen vertaald naar gestructureerde query's om te filteren, selecteren en combineren en ongestructureerde query's die tekstuele eigenschappen rangschikken en vergelijken. Terwijl de gestructureerde bewerkingen in theorie kunnen worden uitgevoerd met behulp van elk SPARQL-systeem, komen bestaande implementaties tekort vanwege hun beperkte ondersteuning bij het verwerken van ongestructureerde gegevens. Bovendien vereist efficiënte vergelijking van ongestructureerde gegevenswaarden (in combinatie met gestructureerde beperkingen) specifieke indexstructuren.



SemaGrow

Alterra, Wageningen UR



Big data analysis

SemaGrow is a project that seeks to help researchers to efficiently question and combine big datasets such as climate data.

The SemaGrow project develops a Linked Data infrastructure that allows access to distributed, heterogeneous and constantly updated large datasets. It aims to tackle this challenge by developing novel algorithms and methods for querying distributed triple stores, scalable and robust semantic indexing algorithms and tools for effective ontology alignment. With the open source software of the SemaGrow Stack, applications can access heterogeneous, distributed triple stores using a single SPARQL endpoint.

To prove its practical value, the SemaGrow Stack is tested in data and knowledge intensive use cases from the agro-environmental domain. They cover aspects like the large heterogeneity of datasets, their often explicit spatial and temporal dimensions resulting in relatively large volumes and their inherent nature of uncertainty, provide challenges which are not usually dealt with. Possible applications vary from access to bibliographical, statistical and multimedia sources by data scientists and educators to querying and integration of distributed Big Data resources for agricultural modelers.

Big data-analyse

Het SemaGrow-project wil onderzoekers helpen grote datasets, zoals klimaatgegevens, efficiënt te bevragen en te combineren.

Het SemaGrow-project ontwikkelt een Linked Data-infrastructuur die toegang biedt tot gedistribueerde, heterogene en voortdurend wijzigende grote datasets. Het project otwikkelt daarvoor nieuwe algoritmen en methoden te ontwikkelen voor het bevragen van verschillende triple stores, schaalbare en robuuste semantisch indexerende algoritmen te ontwikkelen en effectieve tools voor afstemming van ontologieën. Met de opensourcesoftware van de SemaGrow Stack kunnen applicaties heterogene triple stores benaderen via één SPARQL-endpoint.

Om de werking in de praktijk aan te tonen is de SemaGrow-Stack getest met data- en kennisintensieve use-cases uit het agro-milieudomein. Uitdagingen zijn de grote heterogeniteit van datasets in dit domein, hun vaak expliciete ruimtelijke en temporele dimensie, de relatief grote volumes en hun onzekerheid. Mogelijke toepassingen variëren van toegang bieden tot bibliografische, statistische en multimediabronnen door datawetenschappers en opleiders tot het uitvoeren van query's en integratie van data uit gedistribueerde bronnen van big data voor modelleurs in de agrarische sector.

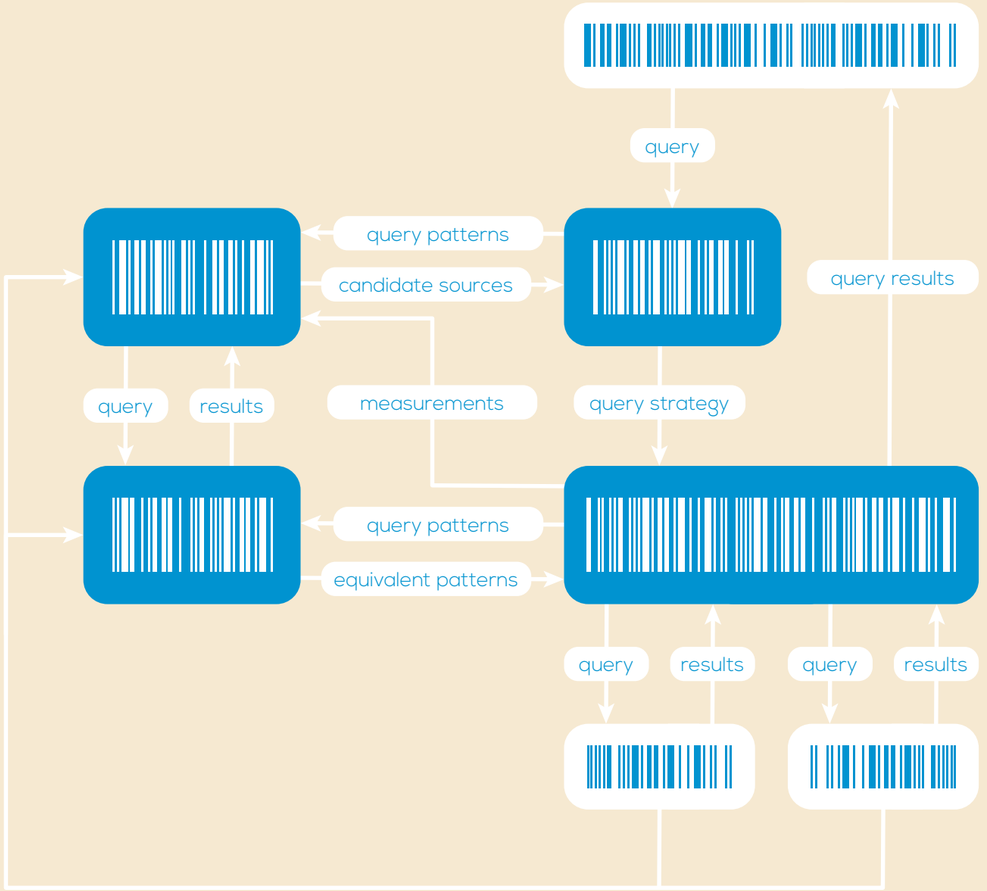
https://github.com/semagrow/semagrow

SemaGrow

Hoe het werkt...

De infrastructuur die SemaGrow heeft ontwikkeld, vereenvoudigt dynamische data-integratie en uitvoering van gedistribueerde queries. Zij is ontworpen voor het werken met grote, gedistribueerde, heterogene, real-time en voortdurend geüpdatete datasets. Hart van de infrastructuur is de SemaGrow-Stack, een federated SPARQL-endpoint waarin meerdere externe SPARQL-endpoints worden geïntegreerd. De SemaGrow Stack gebruikt ontologie-alignment voor harmonisatie van heterogene schema's en kunstmatige intelligentie-methoden om complexe eindpointselecties uit te voeren. De ontologie-alignment en dynamische transformatie van vocabulaires zorgen voor harmonisatie over de uiteenlopende vocabulaires die worden gebruikt om gegevensbronnen in het landbouw- en milieudomein te beschrijven. De endpointselectie gebruikt automatisch gemeten en vastgestelde metagegevens over de inhoud van de verbonden endpoints. Dit levert het gewenste detailniveaudat over het algemeen niet aanwezigis in handmatig aangemaakte metadata. Voor het geval een endpoint niet beschikbaar is, zijn er fall-backmechanismen ingebouwd.

In de architectuur van de SemaGrow Stack wordt data-integratie uitgevoerd door de Query Decomposition-component die query strategieën construeert en doorgeeft aan de Query Manager die daarmee voor de gehele query bepaalt welke patronen voor welk aangesloten endpoint worden gebruikt. De Query Manager zorgt voor uitvoering van de querystrategie en voor het selecteren van alternatieve strategieën wanneer een endpoint niet beschikbaar is. Querystrategieën worden geoptimaliseerd op basis van metadata over (gemeten) beschikbaarheid en responstijd van elk endpoint, het type data dat het levert, de dataschema's en beschikbare mappings tussen verschillende schema's. Deze metadata wordt aangeleverd door de Resource Discovery component die ook de potentiële databronnen suggereert voor een bepaalde query. Resource Discovery werkt met content- en schemametadata. Dit biedt gebruikers de mogelijkheid om nieuwe bronnen te beschrijven en toe te voegen aan de federatie. De Resource Discovery gebruikt onder andere statistische analyse van de performance en betrouwbaarheidsgegevens, gemeten door de Query Manager. De integratie van semantisch heterogene data wordt uitgevoerd door de Query Transformer die gebruik maakt van schemametadata en informatie verkregen via ontology alignment om transformaties uit te kunnen voeren tussen vergelijkbare entiteiten in verschillende schema's.



GVK Online

Ministerie van Veiligheid en Justitie

One glossary for immigration procedures

Right interpretation and understanding of information exchanged between governmental organisations is very important during immigration procedures.

To increase quality of exchanged information and create better understanding between government officials during the immigration process, the Ministry of Safety and Justice developed an interorganisational glossary. This glossary can be consulted via a website and is based on Linked Data technology (RDF, SKOS).

Via the website users are able to quickly search for and hyperlink to definitions. Semantic interoperability between governmental organizations in the immigration process will increase and enhance due to Linked Data technology. The current application will be extended with organisational glossaries. In the future a link can also be made with the glossaries from criminal law. This way a semantic network develops, increasing the semantic interoperability in the information chain.

Eén begrippenlijst voor immigratieprocedures

De juiste interpretatie en begrip van informatie die overheidsorganisaties onderling uitwisselen in een immigratieprocedure is van groot belang.

Om de kwaliteit te verhogen van uitgewisselde informatie en voor beter begrip te zorgen tussen overheidsmedewerkers gedurende het immigratieproces heeft het ministerie van Veiligheid en Justitie een interorganisatorische begrippenlijst opgesteld. Deze begrippenlijst is toegankelijk via een website die gebruik maakt van Linked Data-technologie (RDF, SKOS).

Gebruikers kunnen via de website snel definities opzoeken. Linked data intensiveert en verbetert op deze manier de semantische interoperabiliteit tussen overheidsorganisaties in de immigratieketen. De huidige voorziening wordt nog uitgebreid met organisatiespecifieke begrippenlijsten en in de toekomst kan een relatie worden gelegd met de begrippenlijst van de strafrechtketen. Zo ontstaat een semantisch netwerk en wordt de semantische interoperabiliteit in de keten vergroot.

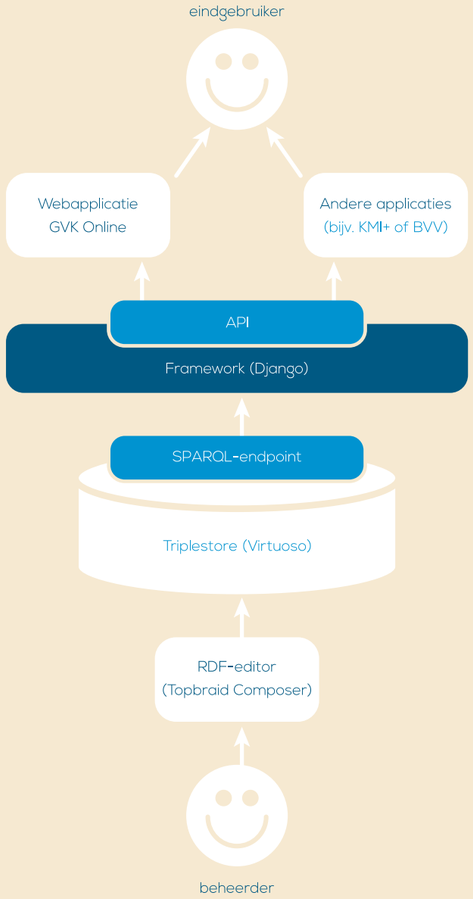
GVK Online

Hoe het werkt...

De belangrijkste modelleringsuitdaging bij het maken van GVK Online was de vraag waar aspecten van de data opgeslagen moesten worden die met de contextafhankelijke representatie ervan te maken hadden. Klassiek voorbeeld: volgorde. Er is uiteindelijk voor gekozen deze representatie-eigenschappen in de applicatielaag onder te brengen en niet in de datalaag (RDF in Virtuoso).

Een andere uitdaging vormde de userexperience en userinterface. Data van het Gegevenswoordenboek Vreemdelingenketen kan een complexe structuur kennen. Ook zijn er heel veel associaties. De userinterface is als een lijst ontworpen, gericht op drie specifieke doelgroepen.

De data en het model van het GVK Online worden beheerd met een RDF-editor (Topbraid Composer). Met deze editor kan men metamodellen/ontologieën beheren. Met de editor kan de data en het model van het GVK Online in verschillende samenhangende RDF-bestanden worden geëxporteerd. Deze bestanden kunnen door de volgende laag (triple store) gemakkelijk worden geïmporteerd.



Om met verschillende personen te kunnen werken aan het beheer worden de data centraal opgeslagen in een repository. Op bepaalde momenten kan een versie worden samengesteld die geschikt is voor publicatie. Deze versie wordt als zodanig opgeslagen in de repository en wordt vanuit daar uitgerold naar de triple store (Virtuoso). De triple store leest de RDF-bestanden vanuit de RDF-editor in en verwerkt deze. Voor het bevragen van triple stores wordt gebruik gemaakt van SPARQL. De data worden middels een gestandaardiseerd SPARQL-endpoint aangeboden aan de volgende laag van de applicatiearchitectuur.

In de applicatiearchitectuur is ervoor gekozen om tussen de webapplicatie (GVK Online-website) en het SPARQL-endpoint een applicatielaag toe te voegen: het framework Django. Deze laag vertaalt de standaard SPARQL-query's naar een standaard API. Met deze API wordt een gedefinieerde set van bevragingen aan het GVK Online ontsloten op een manier die makkelijk in een webapplicatie kan worden verwerkt. Hierdoor hoeft de webapplicatie minder business-logica te bevatten.

Dutch Ships and Sailors

VU Amsterdam

Surfing through maritime history

The Dutch Ships and Sailors dataset allows maritime historians to do historical research in new ways.

Via search and query historians can explore data sets and the links between them. Thanks to an API they can download parts of data for further analysis. The connections between the various datasets make the work of the historians and amateur researchers more effective and efficient.

The Dutch Ships and Sailors datacloud (DSS) is an example of a connected but heterogeneous knowledge graph. The individual data models have been developed in close cooperation with the researchers, which leads to a higher quality and degree of trust. DSS uses PROV-O provenance and Named Graphs for historical requirements for data quality and accountability support. DSS is hosted on a Clio Patria semantic server. An innovative graph-browser has been developed to support the surfing through the knowledge graphs intuitively.

Surfen door de maritieme geschiedenis

De Dutch Ships and Sailors-dataset stelt maritiem historici in staat om op nieuwe manieren historisch onderzoek te doen.

Met behulp van de zoek- en raadpleeg-omgeving kunnen historici datasets en de koppelingen daartussen verkennen. Dankzij een API kunnen zij vervolgens deeldatasets downloaden voor verdere analyse. Juist de verbindingen tussen diverse datasets kunnen het werk van de historici en amateuronderzoekers effectiever en efficiënter maken.

De Dutch Ships and Sailors-datacloud (DSS) is een voorbeeld van een gelinkte maar heterogene knowledge graph. De individuele datamodellen zijn ontwikkeld in nauwe samenwerking met de onderzoekers, wat tot een hogere kwaliteit en graad van vertrouwen leidt. DSS maakt gebruik van PROV-O provenance en Named Graphs om historische vereisten voor datakwaliteit en controleerbaarheid te ondersteunen. DSS wordt gehost op een ClioPatria semantic server. Er is een innovatieve graph-browser ontwikkeld om gebruikers op intuïtievere manier door de kennisgraaf te kunnen laten browsen.

http://2016.semantics.cc/dutch-ships-and-sailors

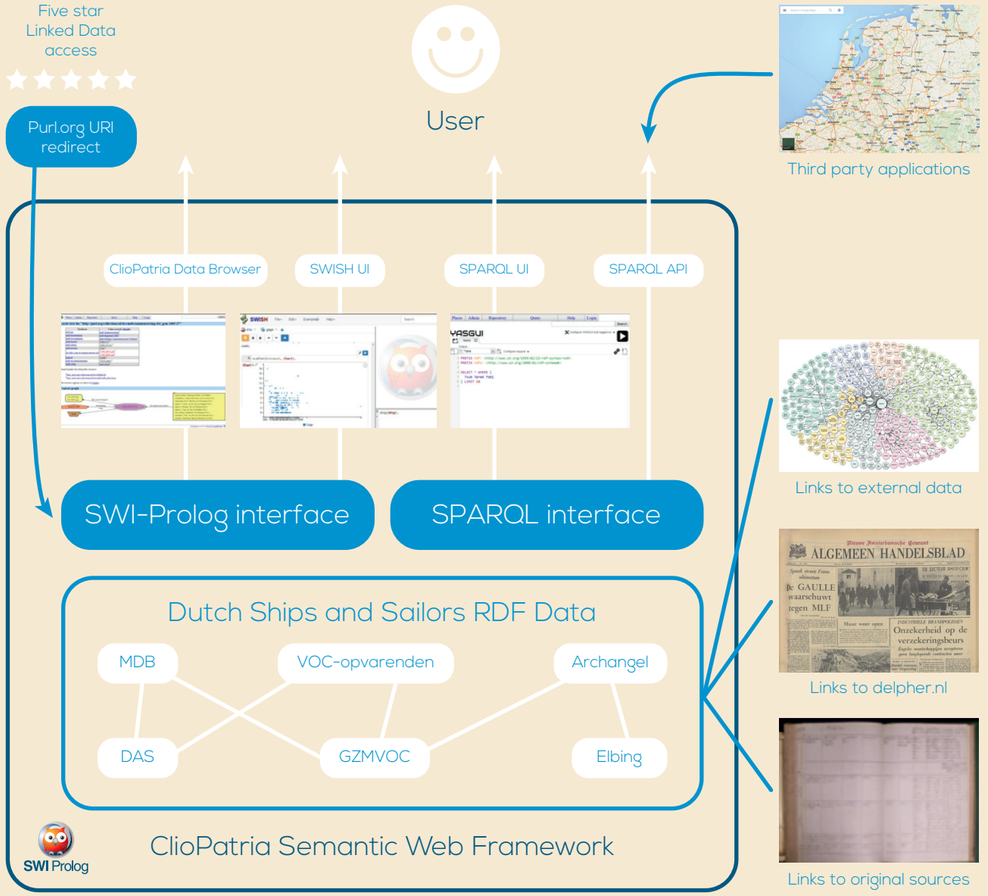
Hoe het werkt...

De Dutch Ships and Sailors-datacloud bestaat uit meerdere maritiem-historische datasets. Deze datasets zijn geconverteerd naar het RDF. Ieder van deze datasets gebruikt zijn eigen datamodel om ervoor te zorgen dat de kennis zo specifiek mogelijk gerepresenteerd is. Elke dataset komt terecht in een RDF named graph, die weer bestaat uit vele RDF triples. Om de herkomst en de status (de zogenaamde provenance) van de verschillende named graphs te administreren is de PROV-O-standaard gebruikt. Door gebruik te maken van deze standaard is terug te vinden waar welke informatie vandaan komt. Dit is een must voor het gebruik van de data voor bijvoorbeeld historisch onderzoek. Behalve named graphs voor de individuele datasets zijn er ook graphs voor onderlinge links, externe links (waaronder links naar oude krantenartikelen gehost door delpher.nl) en links naar online beschikbare scans van de originele archiefstukken.

Alle named graphs samen vormen de DSS-datacloud. Deze RDF-bestanden zijn live geladen in een ClioPatria semantic server. Op dit moment draaien er twee versies live: een stabiele op http://dutchshipsandsailors.nl/data en een ontwikkelserver op http://semanticweb.cs.vu.nl/dss. Het ClioPatria framework bevat behalve een triplestore ook een aantal API's en userinterfaces. Zo is er een SPARQL-API en een tweetal SPARQL-userinterfaces, waarmee eindgebruikers de data kunnen bevragen. De SPARQL-API kan gebruikt worden door externe applicaties. Een voorbeeld is een applicatie die geografische gegevens visualiseert (http://entjes.nl/jeroen/thesis). De ClioPatria-webinterface geeft ook de mogelijkheid om full-tekst te zoeken door de gehele datacloud en om met behulp van een grafische browser het kennisnetwerk te doorlopen. Tenslotte is er ook de mogelijkheid om via de SWISH-webinterface de data te visualiseren en te bevragen met behulp van SWI-Prolog. Via een collaboratieve omgeving kunnen deze query's en visualisaties gedeeld worden.

ClioPatria geeft ook de mogelijkheid om via content negotiation de data als vijf-sterren Linked Data beschikbaar te maken. Zodra een RDF-resource wordt opgevraagd via HTTP, kan op basis van het typerequest, ofwel de webinterface getoond worden, ofwel de ruwe RDF-data geretourneerd worden. Dit laatste kan dan in verschillende formaten, waaronder RDF/XML, Turtle of JSON.

Door deze toepassing zijn enkele technische uitdagingen getackeld. Ten eerste was er de vraag hoe de zeer verschillende bronnen aan elkaar gekoppeld konden worden zonder dat belangrijke informatie in de afzonderlijke bronnen verloren ging. Hiervoor is een methode gebruikt waarbij in de initiële conversie niet of nauwelijks aan normalisatie van gegevens gedaan is. Dit leidt tot zeer uiteenlopende datamodellen. Door gebruik te maken van schema-mappings kunnen de verschillende databronnen weer wel via gemeenschappelijke Classes en Properties bevraagd worden. Een andere uitdaging is hoe deze data toegankelijk gemaakt konden worden voor onderzoekers en leken. Hiertoe is een aantal interfaces ontwikkeld, waaronder een waarin onderzoekers beperkte tot zeer uitgebreide query's live kunnen ontwikkelen, testen, van commentaar voorzien en delen.



Concept library for the Built environment

BIM Loket

Concepts on which you can build

Miscommunication in the construction industry is a head of expenditure. Due to the lack of a clear common language, information can get lost or misinterpreted in the digital transfer from one system to another.

By defining a common language, the concept library for the built environment (CB-NL) aims for better interoperability among the stakeholders of the construction industry. CB-NL is a digital dictionary/taxonomy based on the semantic definition of generic and re-usable concepts (types). These concepts apply to physical objects, functional spaces, and properties, which are related to each other. It is usable in the whole lifecycle and covers the building subsectors: utilities, housing, civil sector etcetera.

The concept library has two main functions. First of all it can be used by organisations to make their own ObjectTypeLibrary, based on the content of CB-NL. Second it is a linking mechanism between different context libraries (including spatial data libraries), which are mapped to CB-NL. The concept library is free to use: it is an open standard, mandated by public clients.

The concept library is based on the Ontology Web Language (OWL). It is a formal OWL specification, not software or a server. It can be downloaded as OWL/RDF file (http://api.cbnl.org/SPARQL/CBNL/statements) and be integrated in other context or applications. The content of CB-NL is offered in a triplestore that can be queried using SPARQL. Its content is available in different formats, as many users are not yet comfortable with OWL/RDF. The content of the library can be viewed with a viewer based on a translation of OWL to XML http://viewer.cbnl.org.

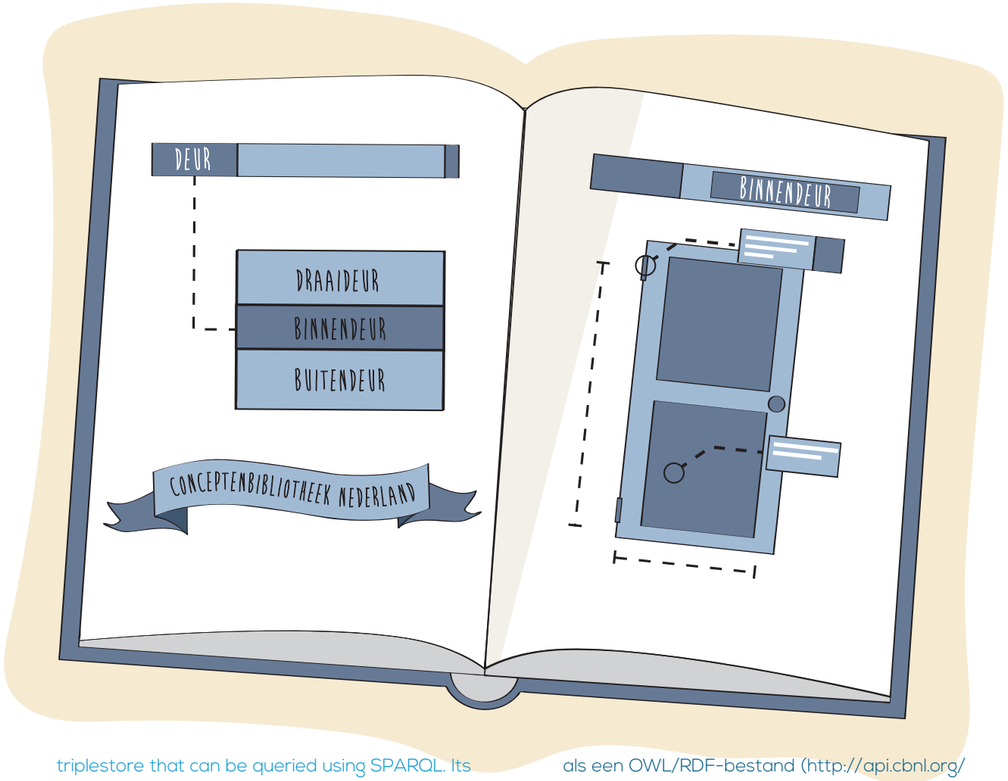
Concepten waarop je kunt bouwen

Digitale miscommunicatie is een kostenpost in de bouw. Omdat er geen gedeelde taal is, kan informatie verloren gaan of verkeerd worden geïnterpreteerd als deze van het ene digitale systeem overgaat naar het andere.

Door deze gemeenschappelijke taal op te stellen, streeft de conceptenbibliotheek voor de gebouwde omgeving (CB-NL) naar betere interoperabiliteit in de bouw. CB-NL is een digitaal woordenboek/taxonomie gebaseerd op de semantische definitie van algemene en herbruikbare concepten (typen). Deze concepten gaan over fysieke objecten, functionele ruimtes en eigenschappen die aan elkaar gerelateerd zijn. Het principe is bruikbaar in de hele levenscyclus van de bouw en omvat deelsectoren als utiliteitsbouw, woningbouw, weg- en waterbouw.

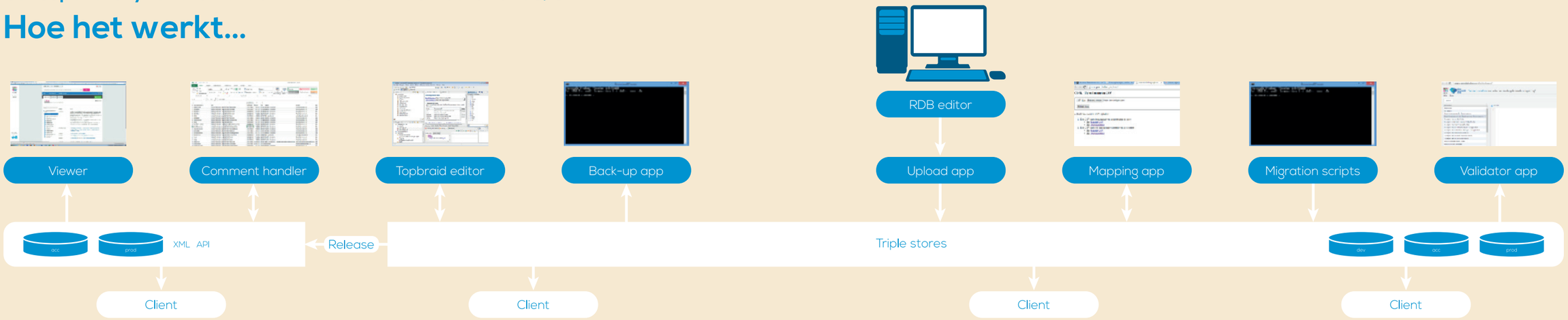
De conceptenbibliotheek heeft twee hoofdfuncties. In de eerste plaats kunnen organisaties de bibliotheek gebruiken om hun eigen ObjectTypeLibrary te maken, gebaseerd op de inhoud van de CB-NL. In de tweede plaats verbindt het verschillende contextbibliotheken (waaronder geo-informatie-bibliotheken), die mappings op CB-NL hebben. De conceptenbibliotheek is openbaar. Het is een open standaard, gemandateerd door gebruikers bij de overheid.

De CB-NL is gebaseerd op de Ontology Web Language (OWL), het is een formele OWL-specificatie. Geen software of server. Het kan worden gedownload als een OWL/RDF-bestand (http://api.cbnl.org/SPARQL/CBNL/statements) en geïntegreerd in andere contexten of applicaties. De inhoud van CB-NL wordt aangeboden in een triplestore die kan worden benaderd met gebruik van SPARQL. De inhoud (of delen ervan) is zonder informatieverlies in diverse formaten opvraagbaar, omdat veel gebruikers nog niet bekend zijn met OWL/RDF. De inhoud van CB-NL is in te zien met een viewer gebaseerd op een vertaling van OWL- naar XML-format: http://viewer.cbnl.org.



http://wiki.cbnl.org

Hoe het werkt...



De Concepten Bibliotheek Nederland (CB-NL) beschrijft concepten in de bouw met Linked Open Data-technologie en drukt dit uit in OWL. Deze aanpak nodigt uit om informatieobjecten in de bouw te relateren naar de CB-NL-concepten, waardoor deze objecten eenduidig getypeerd zijn. Linked open data en ook het OWL-vocabulaire zijn uitbreidbaar, waardoor ook anders ingerichte conceptbibliotheken gerelateerd kunnen worden. Er ontstaat zo een 'tolk'-functie.
CB-NL is de centrale plek waar gemeenschappelijke concepten worden beheerd (hub-and-spoke).

Technische infrastructuur

Het CB-NL-team beheert de CB-NL-ontologieën in twee Sesame triple stores: één voor de bewerking (DEV, ontwikkeling en ACC, acceptatie) en één voor productie (PROD). Via named graphs zijn deze gescheiden van elkaar in een Sesame repository.
Het beheer (op DEV) van de ontologieën wordt met name door middel van TopBraid Composer uitgevoerd. Deze werkt via 'remote access for Sesame' op de DEV triple store. Wanneer een 'release' is afgerond wordt deze informatie overgedragen naar de acceptatieomgeving voor review door een klein aantal experts. Bij goedkeuring wordt de ontologie in de productieomgeving geplaatst zodat deze toegankelijk is via het web.

In de opbouw van het CB-NL is gebruik gemaakt van allerlei tools om de ontologie een eerste vulling te geven. Er is gebruik gemaakt van een externe editor van D.O.N. bureau, een webapplicatie op een relationele database. Deze basale aanpak heeft een goede aanzet gegeven tot de feitelijke bewerking en publicatie. De data zijn door een speciale app omgezet naar OWL en geïmporteerd in de DEV-omgeving. Daarnaast is gebruikgemaakt van migratiescripts, o.a. ontwikkeld in XSLT 2.0 en LODRefine.

Naast TopBraid is ten behoeve van de beheerders een 'validator'-app ontwikkeld. Deze geeft toegang tot alle regels die zijn geformuleerd om een integere ontologie te kunnen vrijgeven. Dit is een webapplicatie en draait direct op DEV en ACC. De app is dus niet geïntegreerd met de bewerkingsomgeving. Basis van de validator-app is een set van SPARQL-query's die gezamenlijk de interne consistentie van de ontologie proberen te bewaken.

Een speciaal onderdeel van de content vormen de 'mappings'. Dit zijn koppelingen tussen concepten van externe partijen en de CB-NL via RDFS- en OWL-vocabulaire zoals rdfs:subClassOf of owl:equivalentClass. Externe partijen kunnen in CSV hun mappings aanleveren en uploaden via een webapplicatie waarna deze worden gevalideerd en omgezet in OWL.

Een SPARQL-endpoint is via http://api.cbnl.org/SPARQL beschikbaar voor clientapplicaties, waaronder een aantal showcases (zie de CB-NL-website: http://public.cbnl.org/oplevering-use-cases-cb-nl). Vanuit het project is een gebruiksvriendelijke viewer ontwikkeld die de triple store bij een nieuwe release uitleest en de gegevens voor de gewone gebruiker 'toegankelijk' maakt. Ook biedt deze viewer een eenvoudig discussieplatform aan. De viewer (http://viewer.cbnl.org/) maakt gebruik van een XML/JSON-API die tevens voor andere applicaties toegankelijk is (http://api.cbnl.org/xml/1.0). Het is binnen het project besloten dat deze toegang relevant is voor de acceptatie van de CB-NL, met name voor andere dan LOD-experts. Er wordt dus niet volledig ingezet op SPARQL, maar is er een tweede weg beschikbaar. Ook de XML/JSON-API en viewer zijn beschikbaar in een OTAP-omgeving.

Volunteered Geographic Information using Linked Data

University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC)

Dealing with ambiguity in data from volunteers

The quality and usability of Volunteered Geographic Information is a subject of debate. Data often comes unstructured with unknown accuracy and lacking reliability. To what extent can Linked Open Data help to semantically enrich volunteered geographic information in order to better answer queries in the context of crisis and disaster relief operations?

To answer this question a proof of concept has been constructed. Data produced by the Ushahidi project during the Chilean earthquake of 2011 has been used as an example. Data were converted into the RDF using vocabularies and semantic links were established to relevant LOD entities. The established links to LinkedGeoData entities have helped to overcome ambiguous georeferencing of the data thus allowing a robust spatial dimension to the data. The semantic enrichments also made it possible to access DBpedia entities via spatial relations. As a result, comprehensive queries could be constructed. For instance, it became possible to prioritise the reports based on the density of population or to extract useful information about local amenities, official names, infrastructural objects, etc. In other words, integration of VGI with LOD provides mechanism to access contextual information thus increasing situational awareness.

The work has shown that the LOD cloud can be perceived as a giant informational skeleton. Scattered and disconnected blobs of unstructured data, being attached to this skeleton, acquire an integrated dataspace where standardized methods of data access and manipulation such as SPARQL can be used. However, non-experts require additional assistance in the interaction with SPARQL endpoints. To tackle this, a prototype software SPEX has been used. With this software a user interacts only with graphical objects and experiences immediate feedback from the manipulations.

Omgaan met tweeslachtigheid in data van vrijwilligers

De kwaliteit en bruikbaarheid van door vrijwilligers geleverde geografische informatie is onderwerp van debat. Gegevens zijn vaak ongestructureerd, de nauwkeurigheid is onbekend en betrouwbaarheid onduidelijk. In hoeverre kan Linked Open Data helpen om de data te bevragen voor rampen- en crisisbestrijding?

Om deze vraag te beantwoorden is een proof of concept gemaakt. Daarbij zijn data gebruikt die zijn verzameld in het Ushahidi-project tijdens de Chileense aardbeving van 2011. De gegevens zijn geconverteerd naar RDF gebruikmakend van vocabulaires, en er zijn semantische links gelegd naar relevante LOD-bronnen. De gemaakte links naar Linked GeoData hebben geholpen onduidelijke georeferenties in de data van vrijwilligers te verbeteren. De semantische verrijking maakte het ook mogelijk om DBPedia-informatie te benaderen op basis van locatiegegevens. Hierdoor konden uitgebreide query's worden gemaakt. Het werd bijvoorbeeld mogelijk om meldingen te prioriteren op basis van bevolkingsdichtheid en om er bruikbare informatie uit te destilleren over lokale voorzieningen, officiële benamingen, infrastructurele objecten enzovoort. Oftewel, de integratie van data verzameld door vrijwilligers met Linked Open Data zorgt ervoor dat informatie context krijgt waardoor er meer inzicht ontstaat in de situatie ter plaatse.

Het proof of concept heeft laten zien dat de LOD-cloud als een informatieskelet werkt. Losse stukjes informatie die aan dit skelet worden gekoppeld, worden hierdoor omgevormd tot een geïntegreerde dataset die met behulp van SPARQL kan worden bevraagd. Aandachtspunt is wel dat niet-deskundigen hulp nodig hebben om met SPARQL-endpoints te kunnen werken. Om hiermee om te gaan is de prototypesoftware SPEX gebruikt. Gebruikers zien alleen grafische objecten en ervaren direct feedback van de handelingen die zij uitvoeren.

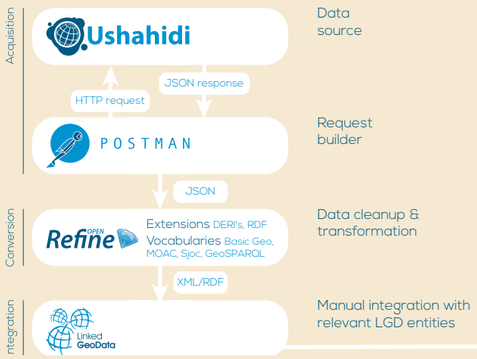
Hoe het werkt...

Voor de bouw van dit proof of concept moest brondata worden verkregen, die moest worden omgezet in RDF door gebruik te maken van ontologieën en ze vervolgens te verbinden met relevante LinkedGeoData-componenten. Dit werd gevolgd door semantische verrijking van de data met gegevens uit LGD en DBpedia. De laatste stap bestond uit het bevragen van de hieruit ontstane dataset om de mogelijkheden van datamanagement te evalueren.

Data acquisitie en conversie

De brondata zijn benaderd met Postman, een extensie op Google's Chrome-browser, die heeft geholpen bij het bouwen en doen van HTTP GET/POST-aanvragen. Elke Ushahidi-deployment heeft een API die rechtstreeks downloaden mogelijk maakt in JSON-format.

Gedownloade JSON-data werden geconverteerd naar RDF-representaties in OpenRefine, een opensourcedesktopapplicatie om data mee op te schonen en transformeren. De basisfunctionaliteit van deze software is uitgebreid met DERI's RDF Refine om export van RDF mogelijk te maken. Verscheidene vocabulaires zijn gebruikt, waaronder het Dublin Core-vocabulaire en de SIOC Core Ontology voor algemene termen als titels, onderwerpen etc. W3C's Basic Geo en OGC's GeoSPARQL-vocabulaires hielpen bij het coderen van ruimtelijke content, terwijl het Management of a Crisis-vocabulaire (MOAC) is gebruikt om de Ushahidi-categorieën om te zetten naar een machine-readable vertaling.

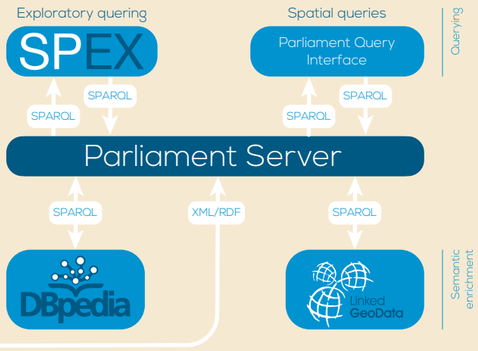


Integratie

Met de ruimtelijke informatie uit de Ushahidi-rapporten werd de basis gelegd voor het leggen van semantische relaties naar de relevante entiteiten in LGD. Tachtig rapporten werden handmatig verrijkt met links naar 95 ruimtelijke objecten die in de rapporten werden benoemd. De resulterende dataset werd geladen in de Parliament triplestore.

Semantische verrijking

Gestructureerde beschrijvingen van de in de rapporten geïdentificeerde ruimtelijke objecten werden teruggehaald van het LGD-endpoint met behulp van SPARQL. De beschrijvingen omvatten objectnamen, classes en geometrie. Die geometrie was van belang omdat het een expliciete ruimtelijke dimensie gaf aan de Ushahidi-data. Hierdoor werd het ook mogelijk om DBpedia-data te benaderen



via ruimtelijke relaties. In het algemeen heeft vrijwel elk bevolkt gebied dat op Wikipedia is beschreven een geografische verwijzing naar een specifieke plek op aarde. Een artikel over een stad kan dus worden gevonden op basis van de locatie van de stad. DBpedia-data zijn bevraagd om informatie te krijgen over gebieden waar de gerapporteerde incidenten plaatsvonden. Deze informatie omvatte gegevens over de populatie, het gebied en de naam van de burgemeester.

Bevragen

Op de semantisch verrijkte dataset zijn diverse query's afgevuurd om de mogelijkheden te evalueren. Door de gegevens uit Ushahidi-rapporten te vatten in triplets en gebruik te maken van de MOAC-vocabulaire om de Ushahidi-categorieën te representeren, werd het mogelijk om op basis van meerdere criteria te filteren. De objectgeometrie vanuit het LGD maakte het mogelijk om op basis van ruimtelijke informatie aanvullende gegevens te vinden. DBPedia-data maakte het mogelijk gegevens uit de rapporten te prioriteren op basis van bijvoorbeeld bevolkingsdichtheid van getroffen locaties. Deze verbeteringen maakten het mogelijk om complexe, samengestelde vragen op de data af te vuren.
Een query haalde bijvoorbeeld alleen de meldingen van geblokkeerde wegen in een bepaald gebied boven. Een ander voorbeeld was het prioriteren van gerapporteerde watertekorten, gebaseerd op bevolkingsdichtheid en het ophalen van de namen van burgemeesters van nabijgelegen steden. Geen van deze query's kon op de individuele oorspronkelijke bronnen worden gedraaid. Bij het maken van de query's is SPEX gebruikt: een tool die het opstellen van een query ondersteunt, en de bestaande Parliament-queryinterface.

Uitdaging

SPARQL-endpoints van zowel DBpedia als LGD warden gedreven door Virtuoso, die de GeoSPARQL-standaard niet ondersteunt. Dit belemmerde de interoperabiliteit van ruimtelijke bevragingen tussen de Parliament triple store en de Virtuoso Server. Oplossing was om data over alle Chileense steden te downloaden uit DBpedia en deze te uploaden in de Parliament triple store. Vervolgens zijn ruimtelijke relaties gebruikt binnen een systeem dat de GeoSPARQL-standaard ondersteunt.

CERISE-SG

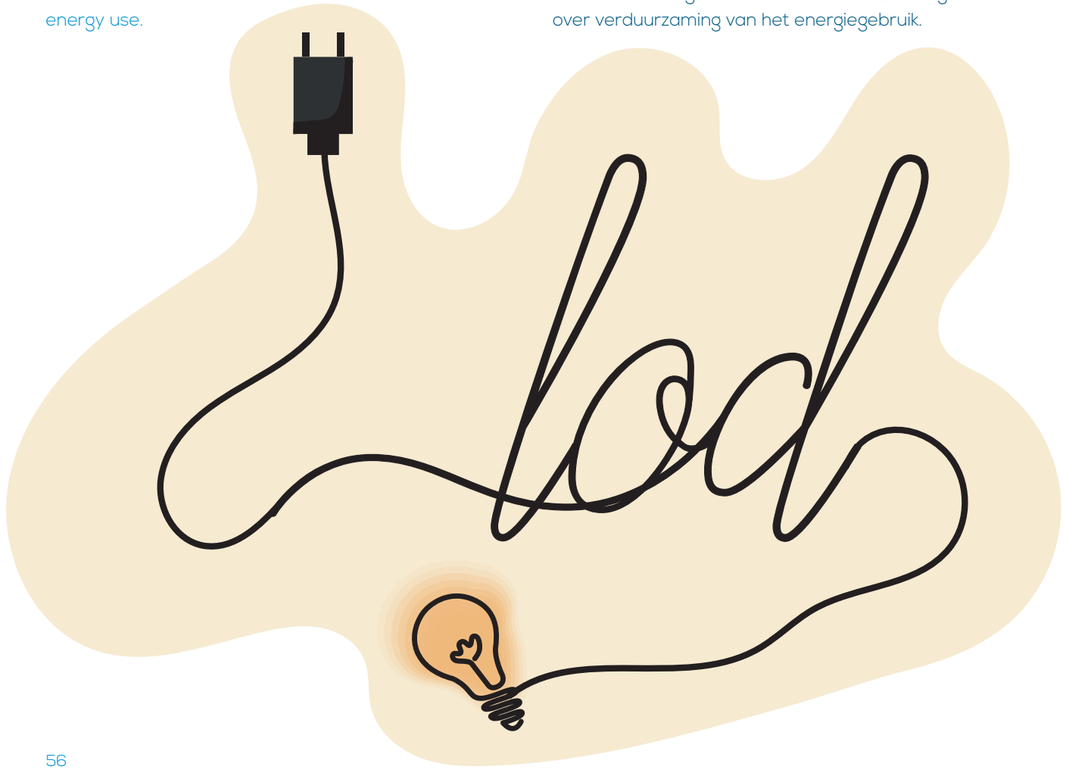
TNO

Makes energy(data) flow

Because energy data are important for many uses and users, wide accessibility is of great importance. Access to this data enables citizens, public administrations, companies and scientists to make better decisions and advise on more sustainable energy use.

Laat energie(data) stromen

Omdat energiedata belangrijk zijn voor velerlei gebruik en gebruikers, is brede toegankelijkheid van groot belang. Toegang tot deze data stelt burgers, overheden, bedrijven en wetenschappers in staat om betere beslissingen te nemen en adviezen te geven over verduurzaming van het energiegebruik.



CERISE-SG shows how Linked Data can be applied to the provision of domain data to the outside world for organisations in the energy domain itself, the public domain and the geographical domain. To the public domain CERISE-SG shows that high quality provision of national data, such as the Dutch Key Registers, has all kinds of useful applications in society. To the geographical domain CERISE-SG shows that spatial data do not necessarily have to be shared as map images, but can be meaningfully applied as raw data.

CERISE-SG shows how geographic data can be published in RDF, and used for GIS procedures. It also shows the power of metadata. CERISE-SG has achieved a great deal in the field of semantic interoperability. Thanks to CERISE-SG, existing models that were not available as Linked Data, made their way to the semantic web – for example, the data model of the Key Register for Adresses and Buildings (BAG) and the Common Information Model (CIM). CERISE-SG also shows that Linked Data is well suited to use in Web applications. Data are made available as JSON and JSON-LD wherever possible.

CERISE-SG laat zien hoe Linked Data kan worden toegepast om domeindata te ontsluiten voor organisaties in zowel het energiedomein zelf als het overheidsdomein en het geografische domein. Voor het overheidsdomein toont CERISE-SG aan dat het hoogwaardig beschikbaar stellen van nationale data, zoals die van de basisregistraties, allerlei nuttige toepassingen kan hebben in de maatschappij. Aan het geografische domein laat CERISE-SG zien dat geografische data niet per se als kaartbeeld gedeeld hoeven te worden, maar ook als ruwe data zinvol kunnen worden toegepast.

CERISE-SG laat zien hoe geografische data gepubliceerd kunnen worden in RDF, en gebruikt kunnen worden voor GIS-procedures. Het toont ook de kracht van metadata op datasetniveau. CERISE-SG heeft veel bereikt op het gebied van semantische interoperabiliteit. Bestaande modellen die nog niet beschikbaar waren als Linked Data, hebben dankzij CERISE-SG hun weg naar het semantisch web gevonden, bijvoorbeeld het datamodel van de BAG en het CIM. CERISE-SG laat ook zien dat Linked Data goed te gebruiken is in webapplicaties. Data zijn zo veel mogelijk beschikbaar als JSON en JSON-LD.

http://www.cerise-project.nl

CERISE-SG

Hoe het werkt...

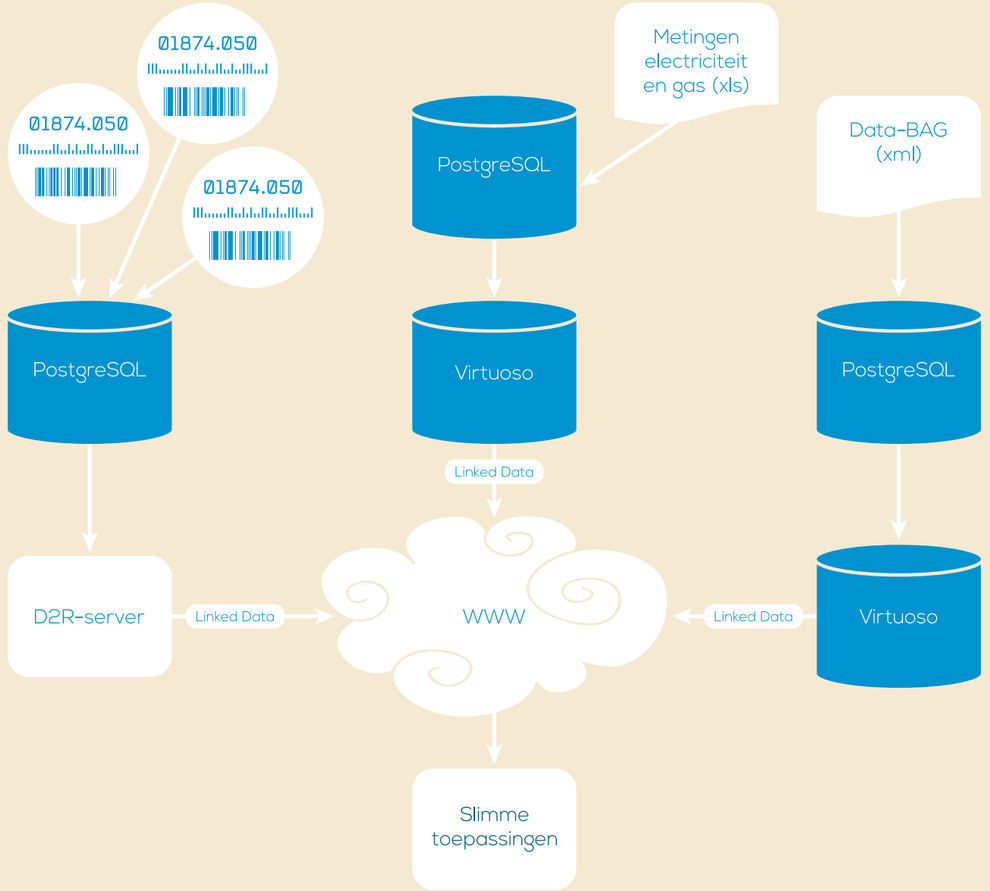
Het doel van CERISE-SG was onderzoek te doen naar het interoperabel maken van verschillende soorten data en verschillende standaarden, om slimme energienetten mogelijk te maken. Het concept Linked Data is toegepast om het wereldwijde web als interoperabiliteitsplatform te gebruiken. In het project zijn gegevens uit het energiedomien en het overheidsdomein, voorzien van locatiegegevens, gepubliceerd als webdata. Om aan te tonen dat dat voldoende is om slimme toepassingen op energiegebied te kunnen maken, zijn enkele demonstratieapplicaties ontwikkeld.

Tijdens het project zijn verschillende datasets (zie de datacatalogus http://lod.geodan.nl/cerisesg/datasets/), verschillende vocabularia en demonstratie-webapplicaties (http://www.cerise-project.nl/) gepubliceerd. Daarbij is geprobeerd om niet alles op dezelfde manier en via dezelfde server te doen, omdat in de werkelijkheid van slimme netten gedistribueerde componenten die van verschillende software gebruik maken interoperabel moeten zijn.

Hieronder een overzicht van software die is gebruikt:

- Voor ontwikkeling van vocabularia en het genereren van HTML-documentatie: Topbraid Composer
- Voor opslag en bewerking van gegevens: postgreSQL en Virtuoso Open Source
- Voor publicatie van data: Virtuoso Open Source, D2R server, ontop [http://ontop.inf.unibz.it/]
- Voor validatie van RDF: Apache Jena
- Voor publicatie van vocabulaires: Apache HTTPD
- Bij ontwikkeling van webapplicaties: jsonld.js, polymer, d3.js

Hoewel de grootste uitdagingen niet technisch van aard waren, kan een gebrek aan toepasbare semantiek bij publicatie van data als Linked Data wel als technisch probleem worden aangemerkt. Dit gebrek aan semantiek trad op bij alle drie de domeinen waarop de focus lag: energie, overheid en geografie. Voor het energiedomein en het overheidsdomein bestond de oplossing uit het zelf ontwikkelen en publiceren van semantiek, een vocabulaire op basis van de BAG (http://lod.geodan.nl/vocab/bag) en een vocabulaire op basis van het CIM (http://ontology.tno.nl/cerise/cim-profile.ttl). Voor geografische data kon worden volstaan met de algemene vocabularia LOCN en GeoSPARQL, hoewel er concessies moesten worden gedaan met betrekking tot het coördinaatsysteem: WGS84 is gebruikt in plaats van ETRS89, dat de voorkeur heeft.



Histograph: geocoding places of the past

Waag Society | Islands of Meaning | Hic Sunt Leones (project: Erfgoed en Locatie)

From 't Haagje to The Hague

Our cultural heritage is a rich source of open data about historical places and events. The availability of such data is growing rapidly, but a geographic search on them is difficult.

Histograph is a geocoder that is being used to standardise and link place names throughout history. For example, different names are being used for the city of The Hague: 's-Gravenhage, Vander Haegen, Haga Comitis, Den Haege, La Haye, De Haach, In den Haige, Schravenhaegen and 't Haagie.

Histograph uses Neo4j and Elasticsearch to expose a web of interlinked toponyms. These are made searchable through an API.

Van 't Haagje naar Den Haag

Ons cultureel erfgoed is een rijke bron van open data over historische plaatsen en gebeurtenissen. De beschikbaarheid van die gegevens groeit snel, maar zijn geografisch moeilijk doorzoekbaar.

Histograph is een geocoder die onder meer gebruikt wordt voor het standaardiseren en koppelen van plaatsnamen door de geschiedenis heen. Zo worden bijvoorbeeld verschillende namen voor de stad Den Haag gebruikt: 's-Gravenhage, Vander Haegen, Haga Comitis, Den Haege, La Haye, De Haach, In den Haige, Schravenhaegen en 't Haagie.

Histograph maakt gebruik van Neo4j en Elasticsearch om een web van onderling met elkaar verbonden toponiemen zichtbaar te maken. Deze zijn gemaakt met behulp van een API.

http://ww.erfgeo.nl

Hoe het werkt...

Met Histograph zijn plaatsen uit het heden en verleden te voorzien van een geometrie en vice versa. Verdwenen straten, gebouwen en dorpen, oude namen, carnavalsnamen, spellingsvarianten, groeiende steden, verplaatste wegen en rivieren zijn allemaal terug te vinden en te tekenen op de kaart. Een set webapplicaties maakt de functionaliteit van Histograph toegankelijk. Deze applicaties zijn nu toegankelijk via het domein www.erfgeo.nl maar in principe is deze tooling in elke webomgeving inpasbaar.

De zoek-API geeft resultaten in GeoJSON-formaat met JSON-LD-context. Resultaten zijn daarmee makkelijk op een kaart te tekenen (bijvoorbeeld met Leaflet) of door de API-output copy/paste in http://geojson.io te laten zien.

Samen met specialisten van de Rijksdienst voor het Cultureel Erfgoed (RCE) en op basis van data uit de primaire use cases is een ontologie ontwikkeld die beschikbaar is op https://api.histograph.io/ontology. De volledige API-specificatie staat op GitHub: https://github.com/histograph/api. Histograph gebruikt graafdatabase Neo4j en Elasticsearch voor indexering van geometrie, tijd en tekst.

Eén van de lastigste vraagstukken was het vinden van het juiste graafmodel om de data in de graafdatabase Neo4J op te slaan. Voor een bestaande techniek als SQL zijn veel 'best practices' beschikbaar en bovendien is het model van zichzelf restrictief en daarmee heel sturend. In een graaf kun je in theorie alles opslaan en relateren, wat het lastig maakt om het systeem performant en efficiënt te laten werken en ook nog inzichtelijk te laten zijn voor gebruikers. Omdat we bovendien niet tevoren alle data kennen en data niet statisch zijn, kan het voorkomen dat men eerst een relatie kent en pas later de node waar die naar verwijst. In Neo4J bestaat die functionaliteit niet en daarom is daarop een abstractie gebouwd.

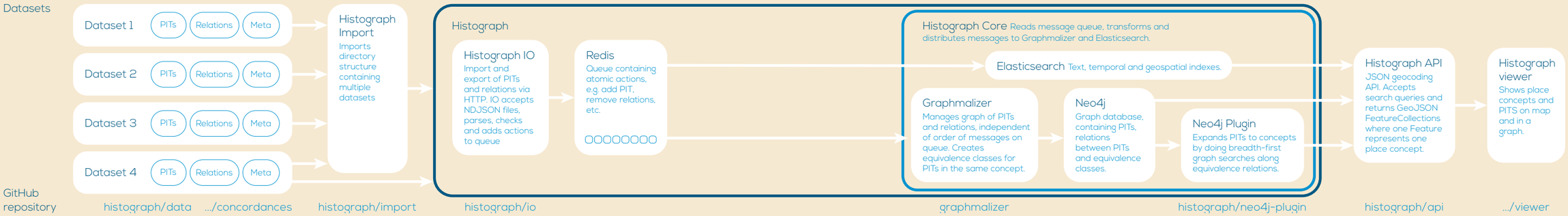
Brondata

De volgende brondata zijn op dit moment opgenomen:

- GeoNames
- Getty Thesaurus of Geographic Names
- Gemeentegeschiedenis
- Kloeke Codes
- poorterboeken
- Verdwenen Dorpen
- VOC Opvarenden 1680-1794
- Historisch Straatnamen Register
- Nationaal Wegenbestand
- Pleiades
- CShapes (nationale grenzen van na de Tweede Wereldoorlog)
- Basisregistraties Adressen en Gebouwen
- Rekeningen Illustere Lieve Vrouwe Broederschap

Resultaten

- http://www.erfgeo.nl
- Geothesaurus vindt elke plaats, bestuurlijke eenheid, straat, gebouw en adres die/dat ooit in Nederland bestaan heeft: http://erfgeo.nl/thesaurus/
- Batchgewijs standaardiseren van plaatsnamen: http://standaardiseren.erfgeo.nl
- Geometrie tekenen en opslaan als geoJSON: http://www.erfgeo.nl/tools/histodraw.html
- Ontwikkelaars die zelf Histograph willen installeren kunnen terecht op http://www.histograph.io



Colofon

Blijf op de hoogte:

- **platformlinkeddata.nl**
- **LinkedIN-groep LOD Nederland**
- **Twitter: #lodnl**
- **Nieuwsbrieven: http://bit.ly/1SFATPh**

Bekijk alle inzendingen voor de Europese prijs voor beste Linked Data-toepassingen op http://2015.semantics.cc/map

Redactie Platform Linked Data Nederland, Yvonne Verdonk | Geonovum
Illustraties Illies, Zwolle
Vormgeving Remwerk, Amersfoort

ISBN 978-90-365-4137-4

Lijst met afkortingen

API	Application Programming Interface
BAG	Basisregistraties Adressen en Gebouwen
CSV	Comma-Separated Values
HDT	Header Dictionary Triples
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
IMS	Identity Mapping Service
IRI	International Resource Identifier
IRS	Identity Resolution Service
JSON	JavaScript Object Notation
LDC	Linked Data Cache
LDF	Linked Data Fragments
LGD	LinkedGeoData
LOD	Linked Open Data
OpenPHACTS	Open PHArmaCological Triple Store
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol And RDF Query Language
SPEX	SPatio-temporal content EXplorer
XML	EXtensible Markup Language
XSLT	EXtensible Stylesheet Language Transformations

Platform Linked
Data Nederland

PlatformLinkedData.nl