**///** OSLO LDES: Thematische werkgroep 1

**////////////////////////////////////////////////////////////////////////////////////////////////////////**

Datum: 20/01/2022

Locatie:  Teams meeting (virtueel)

Moderators: Thijs Hegge, Arne Van Der Stuyft, Pieter Colpaert, Annelies De Craene, Dwight Van Lancker, Philippe Michiels

**////////////////////////////////////////////////////////////////////////////////////////////////////////**

**Aanwezigen**

* Digitaal Vlaanderen
  + Thijs Hegge
  + Annelies De Craene
  + Geert Thijs
  + Arne Van Der Stuyft
  + Filip Borloo
  + Dimitri Hertoghs
  + Nico Schaetsaert
  + Oliver May
  + Dwight Van Lancker
  + Bert Van Nuffelen
  + Simon Claus
  + Veerle Beyaert
  + Adriaan Walpot
  + Geraldine Nolf
* IMEC
  + Pieter Colpaert
  + Philippe Michiels
* IDLab - Universiteit Gent
  + Brecht Van de Vyvere
  + Ruben Verborgh
* District09
  + Maarten Segers
* Vlaams Instituut Voor de Zee (VLIZ)
  + Milan Lamote
* De Watergroep
  + Guy Dillen
* Agentschap Wegen en Verkeer
  + Joris Hoogeboom
* Atos
  + Roland Schmidt
* Konsolidate
  + Christophe Cop
* Esri Belux
  + Gert Bergers

MyCSN NV

* + Jan Geukens
* NSX
  + Frans Verstreken
  + Philip Huysmans
* Cronos
  + Tom Michiels
* ITS Belgium
  + Peter Van der Perre
* Digipolis
  + Michael Meersman
* Zonder toegewezen organisatie
  + Hayden Sutherland
  + Seth Vanhooland
  + Gert de Tant

**Agenda van de werkgroep**

|  |  |
| --- | --- |
| 09u00 – 09u10 | **Welkom & introductie** |
| 09u10 - 09u30 | **Overzicht: wat hebben we gedaan in de vorige werkgroep** |
| 09u30 – 09u40 | **Korte herhaling OSLO & LDES** |
| 09u40 – 10u10 | **Bestaande standaard: Semic** |
| 10u30 – 10u40 | **Pauze** |
| 10u40 – 11u50 | **Een eerste poging tot een data model: feedback en brainstorming** |
| 11u50 –  12u | **Volgende stappen** |

1. **Inleiding**

*[We verwijzen naar de slides voor meer informatie.]*

1. **Overzicht: wat hebben we gedaan in de vorige werkgroep?**

*[We verwijzen naar de slides voor meer informatie.]*

Er werd een overzicht gegeven van de activiteiten van de business werkgroep, dewelke georganiseerd werd op 16 december 2021:

* OSLO-introductie
* Introductie tot Linked Data Event Streams (LDES)
* Brainstorm oefeningen:
  + Capteren van de verwachtingen;
  + Oplijsten welke problemen en/of verwachtingen opgelost zouden kunnen worden d.m.v. een LDES;
  + Capteren van terminologie die de deelnemers terug wensen te zien komen in de specificatie.

De uitkomsten uit deze brainstorm oefeningen werden geclusterd ([zie het verslag van de business werkgroep](https://github.com/Informatievlaanderen/OSLOthema-ldes/raw/standaardenregister/reports/Verslag%20business%20werkgroep%20-%2016%20december%202021.docx)). Deze clusters werden gebruikt als rode draad in deel 3 ‘korte herhaling OSLO & LDES’ van deze werkgroep om aan te tonen wat OSLO en Linked Data Event Streams kunnen betekenen voor de deelnemers.

1. **Korte herhaling OSLO & LDES**

*[We verwijzen naar de slides voor meer informatie.]*

In dit deel van de thematische werkgroep werd er gefocust op wat OSLO en Linked Data Event Streams kunnen betekenen voor meer specifieke problemen en verwachtingen die in de business werkgroep werden gecapteerd, namelijk:

* Nood aan een standaard voor het aanbieden van data;
* Nood aan een oplossing voor de problemen omtrent replicatie en synchronisatie;
* Nood aan een oplossing voor de ‘maintenance hell’;
* Nood aan een aanpak voor versiebeheer;
* Nood aan een duidelijke basis-definitie van een Linked Data Event Stream;
* Nood aan community building en kennisdeling.

1. **Bestaande standaard: Semic**

*[We verwijzen naar de slides voor meer informatie.]*

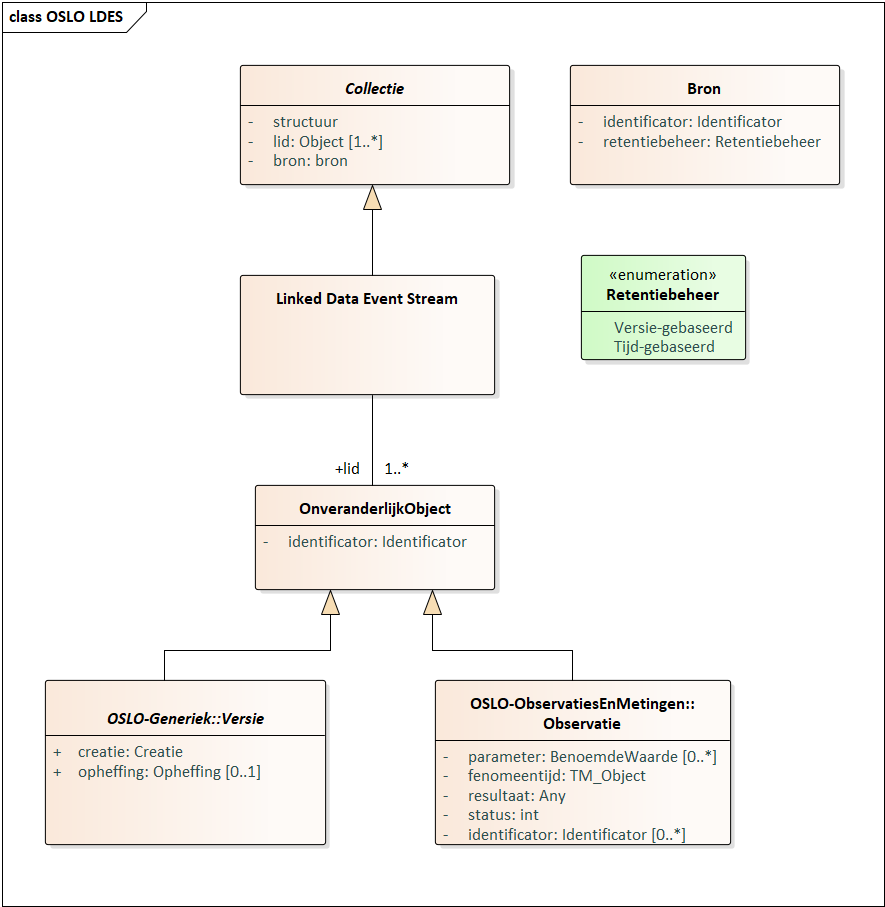
Aangezien we voor onze eerste poging tot een datamodel in deel 5 (zie hieronder) vertrekken van de bestaande [SEMIC specificatie](https://w3id.org/ldes/specification) op Europees niveau, werd deze eerst toegelicht en geïllustreerd.

1. **Een eerste poging tot een data model: feedback en brainstorming**

*[We verwijzen naar de slides voor meer informatie.]*

Tijdens deze fase werd een eerste poging tot een datamodel (zie afbeelding 1 hieronder) voorgesteld aan de deelnemers. Met behulp van de online samenwerkingstool [Mural](https://app.mural.co/t/beadvtc7549/m/beadvtc7549/1642058003282/c4fc00efe50465f570d5149b04acd586f51c6a39?sender=u048a1117151baed084666519) werden de verschillende onderdelen van het model plenair voorgesteld, met hun respectievelijke definitie, waartoe het kernteam een aanzet gedaan had.

Vervolgens werden de deelnemers telkens verzocht om in drie herhalende breakout rooms te discussiëren over het respectievelijk voorgestelde onderdeel van het datamodel. Na elk van de sessies in de breakout rooms werden de discussies en opmerkingen plenair verzameld, overlopen en geconsolideerd. De verschillende onderdelen van het model, alsook de gecapteerde feedback van de deelnemers erop, worden hieronder beschreven.



*Afbeelding 1: Eerste poging tot datamodel*

* 1. **Basisdefinitie linked data event stream**

In het eerste deel van deze oefening werd gefocust op de algemene definitie van een Linked Data Event Stream. Hiertoe had het kernteam de volgende aanzet gedaan (op basis van de SEMIC Sspecificatie):   
  
*“Een* ***Linked Data Event Stream*** *(LDES) is een collectie van onveranderlijke objecten (zoals versie-objecten, sensor-observaties of gearchiveerde representaties).”*  
  
Op bovenstaande definitie werd de volgende feedback gegeven door de deelnemers:

1. Wat is een gearchiveerde representatie juist?  
     
   *Hiermee wordt een oudere versie van een onveranderlijk object bedoeld, wat op zijn beurt als gearchiveerd wordt gepercipieerd omdat er een nieuwe versie is die verwijst naar dat object.*

1. Wat is ‘Bron’ in deze context?   
     
   *Zie deel 2 van deze oefening.*
2. Hoe moet archivering juist werken? Werkt dit met behulp van views of zijn archiefstromen afgeleide datastromen? Moet er dan ook metadata voorzien worden om daarnaar te verwijzen? Hoe gaan we om met de situatie waarin een deel van de Linked Data Event Stream wel al gearchiveerd is en een deel nog niet?

*In OSLO Generiek gebeurt dit door niet enkel het moment op te geven waarop het record is ontstaan, maar ook het moment mee te geven waarop het record beëindigd is, namelijk creatie en opheffing.*

1. Wat als je meerdere servers hebt die dezelfde LDES publiceren maar verschillende retentiebeleiden volgen? Bv. één server die de data maar één dag bijhoudt, en een andere server die de data twee jaar bijhoudt; hoe pakt men dit aan?

*In een metadata portaal moet een volledig overzicht gegeven kunnen worden van alle servers die werken met een originele Linked Data Event Stream. In dat zelfde metadata portaal zou men ook moeten kunnen zien welk retentiebeleid aanwezig is op welke server.   
  
Hier zal dieper op ingegaan worden in een volgende thematische werkgroep, waarin we zullen werken rond het beheer van metadata.*

1. Er is nood aan een minimum aan metadata om te vermijden dat eenzelfde LDES op verschillende manieren wordt ingevuld, wat het moeilijk zou maken om consistent gegevens te verwerken.

*Dit kan door gebruik te maken van andere OSLO standaarden. Voor observaties is dit ‘OSLO Observaties en Metingen’, waarin ook alle metadata zit begrepen. Deze OSLO standaarden zijn gebaseerd op internationale standaarden. LDES heeft het doel om data van alle types bestaande standaarden gemakkelijk te kunnen repliceren en synchroniseren (d.m.v. een interface ervoor zorgt dat alle data afnemers onmiddellijk gaan synchroniseren zodra ze hebben gerepliceerd).*

1. Dient de term ‘Event stream’ mee opgenomen te worden in de naam? Dit zou verwarrend kunnen zijn doordat het het beeld geeft dat het enkel over een metadata-stroom gaat (een stroom die naast de werkelijke data leeft en aangeeft wanneer er iets verandert).   
     
   *Het gaat hier niet enkel over de metadata-stroom, maar dit zou ook wel mogelijk moeten zijn binnen een Linked Data Event Stream. Een metadata-stroom die aangeeft dat er iets verandert in de eigenlijke data en de data afnemer naar die data doorverwijst om de eigenlijke verandering op te halen.*

*Een Linked Data Event Stream is meer dan een metadata-stroom. Het is ook een logboek met de eigenlijke versie-objecten in. De opmerking doelt niet op de verandering van de basisdefinitie, wel op het vermelden van de ruimheid van de definitie wanneer er in de toekomst opleidingen, trainingen etc. zouden georganiseerd worden.*

1. Wat is het verschil tussen dataset, service en collectie?   
     
   *Dit wordt meegenomen naar één van de volgende thematische werkgroepen die zal kaderen rond metadata.*
2. In een Linked Data Event Stream wordt een “shape” gedefinieerd, waaraan alle objecten moeten voldoen. Het is belangrijk dat die “shape” altijd backwards compatible is aangezien er met onveranderlijke objecten gewerkt wordt.

*Aangezien de onveranderlijke objecten in de collectie backwards compatible moeten zijn, dienen alle onveranderlijke objecten in één Linked Data Event Stream volgens een bepaalde shape gepubliceerd te worden (bv. SHACL, SHEX, …). Wanneer de definitie van die shape verandert, dient overwogen te worden om naar een andere Linked Data Event Stream te verwijzen door het feit dat de huidige Linked Data Event Stream niet meer backwards compatible is. Hieruit volgt dat ‘shape’ een cruciaal onderdeel is van een Linked Data Event Stream, en daarom ook zal worden toegevoegd aan de basisdefinitie.*

*Echter blijft de data publisher steeds in controle over wat wel en niet wordt toegelaten in een Linked Data Event Stream.*

*In de specificatie zal duidelijk weergegeven worden hoe een data publisher moet omgaan met de shape, omdat de data publisher per basisgegeven (authentieke bron) de levensloop moet bepalen. Hij/zij moet vastleggen welke wijzigingen worden toegelaten om nog over dezelfde Linked Data Event Stream te kunnen blijven spreken.*

1. Mag een Linked Data Event Stream een mix van objecten van verschillende types zijn? En zo ja, hoe geven we dan de samenhang weer tussen die objecten?

*Bijvoorbeeld: mogen straatnamen en adressen in eenzelfde LDES gepubliceerd worden, en hoe geven we dan de samenhang weer tussen straatnamen en adressen? Per definitie zal een adres altijd een straatnaam hebben, en dus zal een adres automatisch veranderen wanneer een straatnaam verandert. Hierdoor moet er wel een temporele samenhang zijn.   
  
Punt 9 komt neer op volgende vraag: Kunnen we een mix van verschillende type objecten (bv. Straatnamen en adressen) in één LDES publiceren of niet?*

*Momenteel zijn hiervoor verschillende opties mogelijk. Je kan zowel naar het algemeen object verwijzen en geen andere data includeren. Daarnaast kan je naar een versieobject verwijzen zodanig dat je dat elders kan raadplegen. Je kan heel het object includeren (bv. adressen in de LDES van straatnamen). Het probleem hierbij is dat wanneer dit geïncludeerd object wijzigt, dit zal resulteren in een update voor de gehele LDES.*

*[We verwijzen hier naar slide 39 t.e.m. 43 voor de verschillende opties]  
  
Het is aldus aan de ontwikkelaar om hier voor zijn/haar specifieke use case een juiste keuze te maken. Wel moet het duidelijk zijn of het over een versieobject gaat of over een referentieobject. Dit zal duidelijk gemaakt worden in de specificatie.*

1. Hoe geven we de samenhang van de onveranderlijke objecten weer? Wordt er impliciet aangenomen dat alle objecten in de juiste ordering staan?

*In de SEMIC specificatie werd dit zo ruim mogelijk gehouden. Alles wat in een stream komt is op een bepaald moment naar die stream geschreven. Dit tijdsaspect kan als een impliciete ordening gezien worden.*

*In OSLO Generiek is er onderscheid gemaakt tussen een wijziging op terrein en een technische wijziging. We verwijzen naar deel 3 van de oefening.*

1. De bovenstaande basis-definitie is redelijk breed; zal deze nog verfijnd worden?   
     
   *Het is een afweging die we moeten maken tussen enerzijds het dekken van alle concepten en anderzijds het voor een ontwikkelaar duidelijk maken wat hij/zij ervan kan verwachten.*

* 1. **Definitie Collectie, Structuur, Lid, Bron**

In deel twee van de oefening werd dieper ingegaan op de kenmerken van een Linked Data Event Stream, namelijk:

* Een **collectie** is een verzameling van objecten die een bepaalde structuur naleven. Elke collectie wordt gepubliceerd op één of meerdere servers (de bron(nen)). Een collectie volgt een bepaalde vastgelegde structuur, en heeft een aantal (één of meerdere) leden.
* De ‘**structuur**’ van een collectie definieert de vormvereisten van de leden. Het geeft weer welke oude en nieuwe leden zijn en gevalideerd zullen worden door deze structuur. Als een gevolg van de onveranderlijkheid van de objecten, kan deze structuur veranderen, maar moet deze altijd backwards compatibel zijn.
* Het ‘**lid**’ geeft weer welke onveranderlijke objecten deel uitmaken van de collectie (en wordt later gespecificeerd naar welke onveranderlijke objecten deel uitmaken van de LDES).
* Het attribuut **bron** geeft de bron (i.e. server) aan op welke alle leden geraadpleegd kunnen worden.

Op bovenstaande definities werd volgende feedback gegeven door de deelnemers:

1. Wat wordt er bedoeld met ‘server’ in de definitie van een bron (zie 4.)?   
     
   *Dit moet duidelijker gespecificeerd worden. Alles wat ‘server’ zegt zal verwijderd worden uit de definitie.*
2. Kan een collectie meerdere bronnen (i.e. als in views), hebben?

*Een collectie moet meerdere bronnen kunnen hebben. Dit zal zo aangepast worden in het UML model.*

1. Het element van fragmentatie ontbreekt. Men zou moeten kunnen duiden welk deel van de fragmentatie in welke bron wordt bijgehouden.   
     
   *Dit zal in de metadata weergegeven worden.*
2. Zijn er nog andere retentiebeleiden mogelijk naast deze van versie en tijd?   
     
   *Hier werd een suggestieve codelijst voorzien. Deze kan uitgebreid worden.*
3. In welke taal zal de metadata geschreven worden?   
     
   *Hier zal het kernteam een antwoord op bieden tegen de tweede thematische werkgroep.*
4. Als je werkt met bv. sensordata, kan men aan de data een parameter toevoegen van de frequentie waarmee deze data binnenkomt?

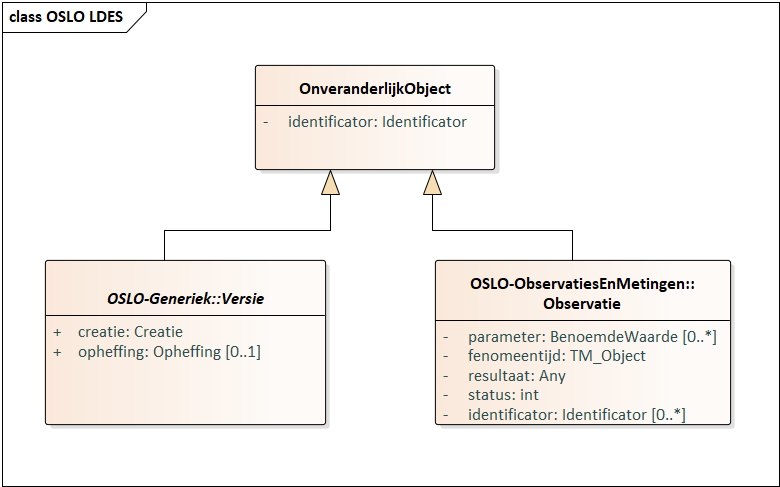
*Dit zal mee worden opgenomen in data lineage (hoe verhouden de data zich tot elkaar). Hierin zal er beschreven worden dat een Linked Data Event Stream of een collectie een relatie (bv. een afgeleide dataset) heeft tot één of meerdere andere Linked Data Event Stream(s).  
  
Anderzijds kan er gekeken worden naar de caching van je HTTP pagina’s en de frequentie waarmee deze gebeurt.   
  
Updatefrequentie is een onderdeel van de metadata DCAT en kan daar ook in opgenomen worden.*

1. De naam bron is ongelukkig gekozen. Bron duidt op de plaats waar de leden van de collectie kunnen gevonden worden.

*Dit zal aangepast worden tegen thematische werkgroep 2. Binnen het kernteam zal bekeken worden of een vertaling van de term ‘view’ nodig is.*

1. Worden verschillende views op dezelfde collectie allemaal apart beschreven in de metadata?   
     
   *Dit zal besproken worden in de werkgroep rond metadata. Dit bestaat binnen DCAT 3.0. Waarnaar er op dat moment zal verwezen worden.* 
   1. **Definitie Onveranderlijk Object**

In deel drie van de oefeningen werd aan de deelnemers gevraagd om feedback te vormen inzake het onveranderlijk object en de types van onveranderlijke objecten. Aan de deelnemers werden twee types van onveranderlijke objecten voorgesteld, namelijk versie en observatie (zie afbeelding 2).



*Afbeelding 2: onveranderlijke object en zijn types*

Volgende feedback werd gecapteerd:

1. Wat is een identificator en is die verplicht bij een onveranderlijk object?  
     
   *De identificator zou niet verplicht moeten zijn. De multipliciteit zal aangepast worden naar 0…\*.*
2. Maken we timestamp verplicht op een onveranderlijk object en welk soort timestamp (wanneer het op de event stream is gekomen of wanneer het werkelijk is gebeurd)?

*Dit is opgelost in versie en observatie in de respectievelijke OSLO applicatieprofielen.   
De vraag blijft wel met welke property beide zullen weergegeven worden. Dit zal toegevoegd worden aan de klasse onveranderlijk object tegen thematische werkgroep 2.*

1. Als er geen timestamp wordt meegegeven, kan je (momenteel) geen ordening terugvinden tussen de objecten van een LDES.

*Dit kan perfect, maar houdt in de meeste use cases dan wel niet al te veel steek. Als we dan vertrouwen op specialisatie van onveranderlijk object, dan dient de vraag gesteld te worden het nut van die klasse nog is in deze context.*

*De types van een onveranderlijk object zullen gekozen worden op een manier dat het verleden niet kan gewijzigd worden (bv. Een adres kan wijzigen, maar een versie van een adres niet).   
  
Dit zal bekeken worden tegen thematische werkgroep 2.*

1. Hoe geef je aan van welk object het een versie is?   
     
   *In OSLO Generiek staat ‘isTijdsSpecialisatieVan’* *om dit aan te geven. Hierin kan je dan de URI weergeven.*
2. Hoe zullen de identificator ingevuld worden voor verschillende soorten objecten?   
    *Structuur en naming conventions hierrond zijn nodig. Linked Data werkt met URI’s als identifiers. Dat is niet voor alle data het geval, waardoor we een extra veld ‘identifier’ toevoegen om dit expliciet weer te geven zodanig dat iedere gebruiker hiermee om kan gaan.*
3. Waarop moet een gebruiker/afnemer synchroniseren?

*Bijvoorbeeld voor een adressenregister is een tijdsweergave heel belangrijk: Wanneer een data afnemer gaat kijken of er wijzigingen zijn gebeurd, moet hij/zij de huidige situatie kunnen vergelijken met een voorgaande situatie.*

*Zo dient er use case per use case bekeken te worden wat er belangrijk is voor een data afnemer om te kunnen synchroniseren en dit kan als identifier gebruikt worden.  
  
Soms zal het voorkomen dat je dubbele identifiers hebt, een URI en een ID (bv. rijksregisternummer). Hier dient ook een mapping voorzien te worden.*

1. **Overige opmerkingen en discussies**

1. Versiebeheer is in de SEMIC specificatie gebaseerd op dcterms:isVersionOf. Echter bestaat er in OSLO reeds materiaal over versiebeheer, namelijk OSLO Generiek dewelke gebaseerd is op de provenance ontologie.

*Antwoord: dcterms:IsVersionOf werd gebruikt ter illustratie. In een volgende werkgroep zal versiebeheer meer in detail besproken worden, waarin ook beslissingen genomen kunnen worden over vragen zoals bv. welke ontologie als basis gebruikt zal worden. Wel wordt hergebruik geprefereerd binnen de OSLO methodologie.*

1. **Volgende stappen**

Indien u graag zou willen deelnemen aan één van de aankomende werkgroepen, kan u via de volgende [link](https://overheid.vlaanderen.be/opleiding/oslo-ldes) een overzicht van de sessies terugvinden en u ook zo inschrijven. De ‘thematische werkgroep 2 OSLO LDES’ is de eerstvolgende werkgroep voor het OSLO-traject en zal plaatsvinden op 17 februari 2022 om 9u00 via Microsoft Teams waarvan de link wordt doorgestuurd naar de deelnemers.

Inschrijven voor de volgende werkgroep (17 februari 2022) kan via onderstaande link: <https://overheid.vlaanderen.be/informatie-vlaanderen/agenda/thematische-werkgroep-2-oslo-ldes>