De BAG in Linked Data

Hoe brouw je 600 miljoen triples

Spreker: Richard Nagelmaeker

ORDINA

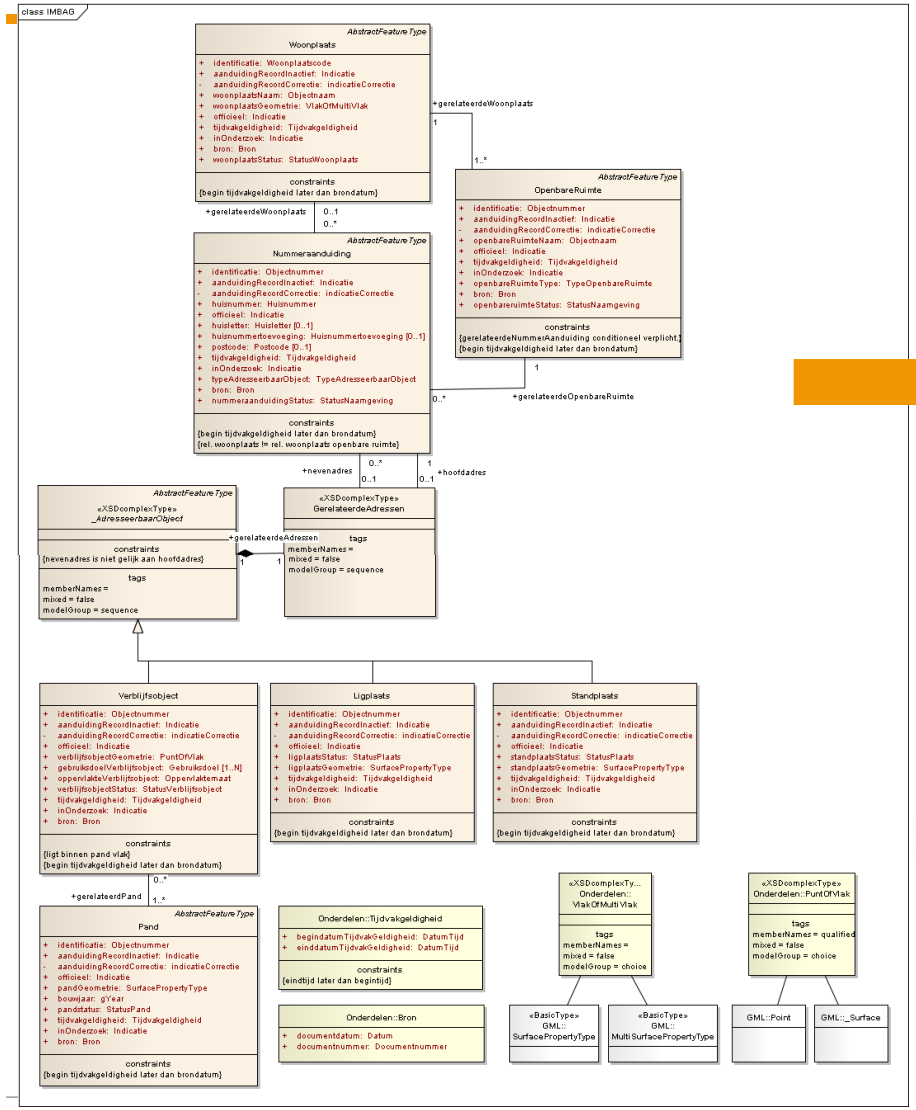
Introductie

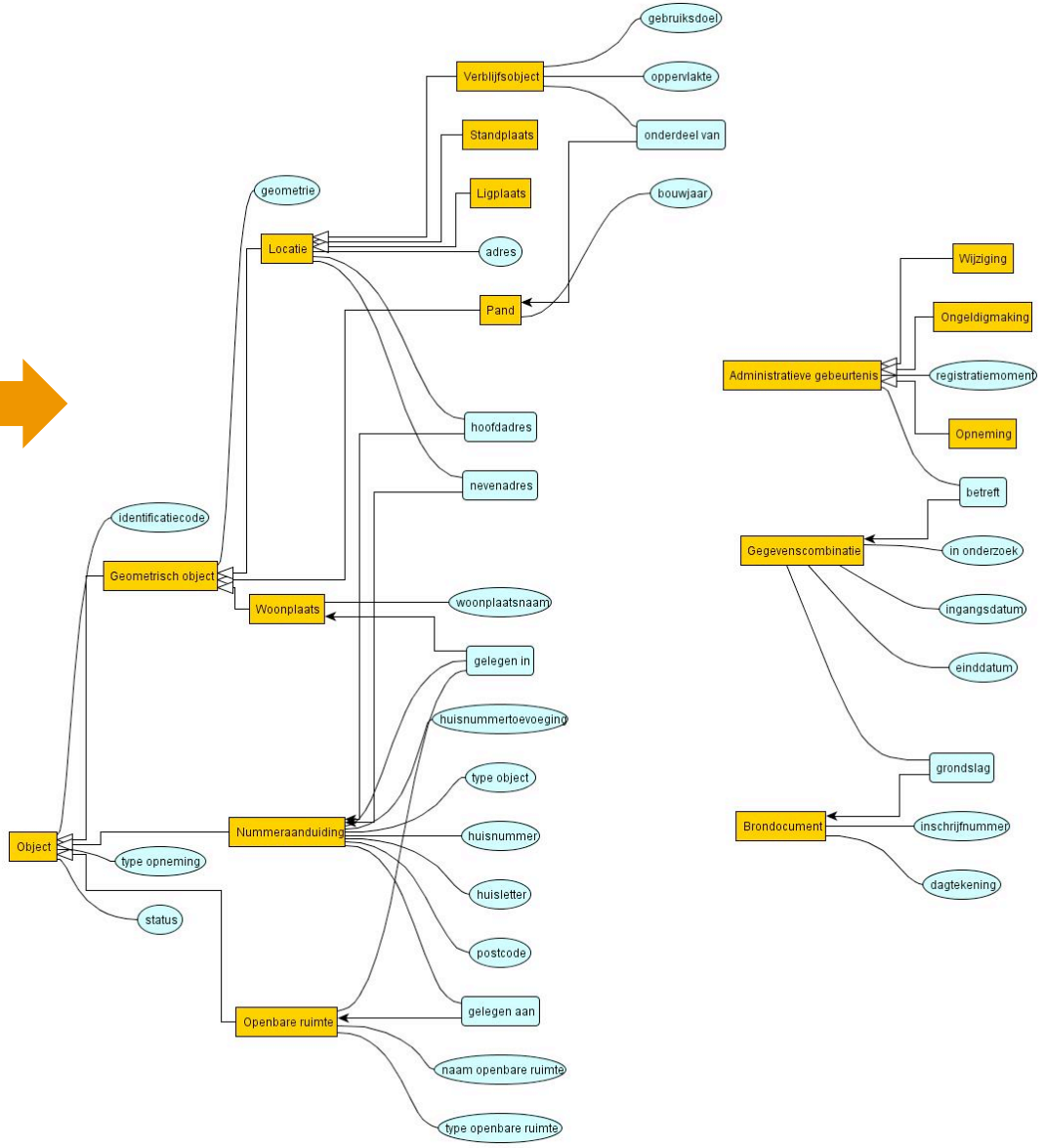
- Richard Nagelmaeker
- Architect Interoperabiliteit – bij RT&I

- Ambitie : Een informatievoorziening die organisaties, medewerkers en klanten daadwerkelijk ondersteund.

De Strijd

Van Relationele Database naar Triples (BAG)





Aanpak

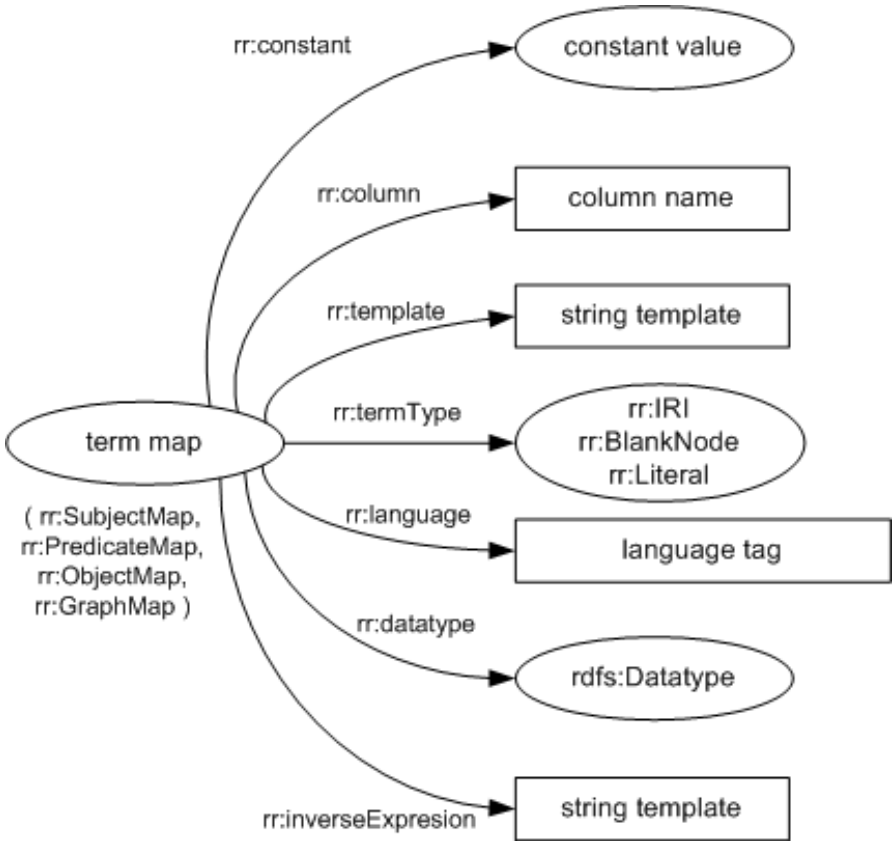
- Bag downloaden via Link van Marcel van Mackelenbergh (Belastingdienst)
- Bag XML files in Postgresql database stoppen, met NLExtract
- Extracten vanuit DB naar triples met R2RML → later met SML
- Triplebestanden opdelen, zodat je meerdere bestanden tegelijk kunt uploaden
- Triplebestanden uploaden naar Virtuoso met rdf_loader_run() vanaf commandline (1 per core starten)
- nalopen met rdfconvert van gewijgerde bestanden (output naar /dev/null)

- Totaal ruim 600 miljoen triples

R2RML – RDB to RDF Mapping Language

- R2RML is een W3C standaard
- R2RML gaat uit van 1 Subject per tabel of SQL query
- Je krijgt dus configuratie set per Subject.
- R2RML wordt geconfigureerd met TTL bestand (configuratie in triples)
- Per Subject kun je Predicates en Objects aangeven (PredicateObjectMap)
- Meerdere Subjects? Je kunt Subject van één configuratie set als object aanhalen in een andere (Subject) configuratieset
- Named Graphs binnen R2RML mogelijk alleen:
  - Per Subject en Per Predicate Object Map de Named Graph aangeven
  - Meeste tools ondersteunen geen TriG of TriX of N-Quads
- Goede documentatie (R2RML standaard W3C) – Raport over tool ondersteuning bied aardig wat voorbeelden voor configuratie van R2RML

R2RML – RDB to RDF Mapping Language



SML – Sparqlify Mapping Language

- SML is proprietary het is een specifieke taal voor Sparqlify
- SML en R2RML werken ongeveer hetzelfde
- SML gaat echter uit van een set aan triples, waarvan de variabelen gevuld worden vanuit een tabel of SQL query
- Meerdere subjects op één SQL query is eenvoudig te realiseren
- Ondersteund Named Graphs
- Je configureert dus voor 1 tabel in één keer

Tools - Vergelijk

	DB2Triples	Sparqlify
R2RML	Ja	(Wordt aan gewerkt)
SML	Nee	Ja
Ondersteuning Named Graphs	Ja, maar schrijft niet weg in Trig of Trix???	Ja
Bestandsformaten	TTL, RDF/XML	N-Triples, N-Quads
Performance	1.000 Triples / s	10.000 Triples /S

Leerpunten

- Kies tooling voordat je DB kiest
- Probeer waar mogelijk de database per tabel te vertripellen. Laat RDF de gegevens (middels foreign keys) maar verbinden (dit gaat vanzelf)
- De triple store brengt alles wel bij elkaar
- N-Quads is te prefereren als output formaat (ivm. snelheid en eenvoud)
- Split grote bestanden op zodat je er meer te gelijk (parallel) in de triple-store kunt laden (één van de voordelen van N-Quads is dat dit gemakkelijk kan)
- Doorloop eerst het gehele proces van A tot Z met een kleine gegevensset (10000 triples is meer dan genoeg). Ga daarna pas met de volledige omvang van de bron aan de slag

Samen innoveren aan een duurzame digitale wereld

ORDINA