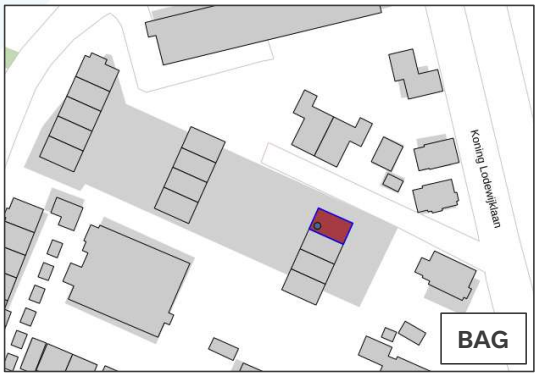
Data Lineage in the IMX Orchestration of Federated Data Sources

Auteur Pano Maria

Datum Wednesday, September 6, 2023

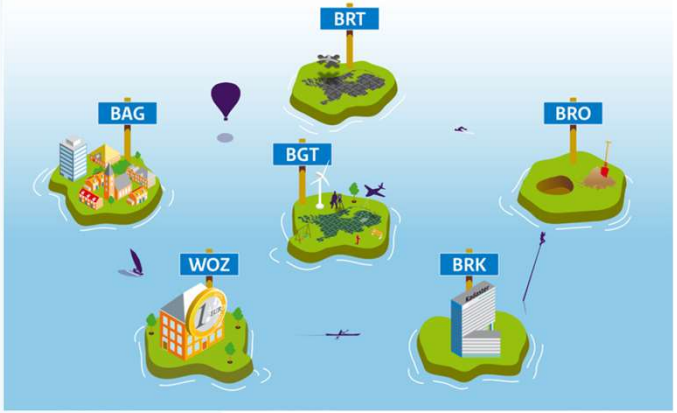
Neptunusstraat, Apeldoorn, June 22 2019

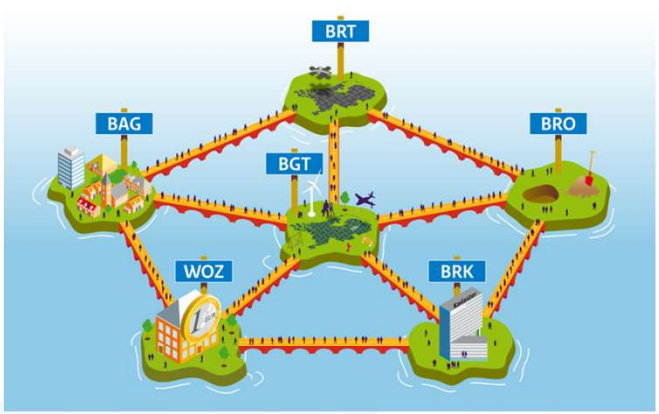
Geo base registries: Different views on reality



Neptunusstraat, Apeldoorn, June 22 2019

From islands to integrated landscape





How to achieve coherence?

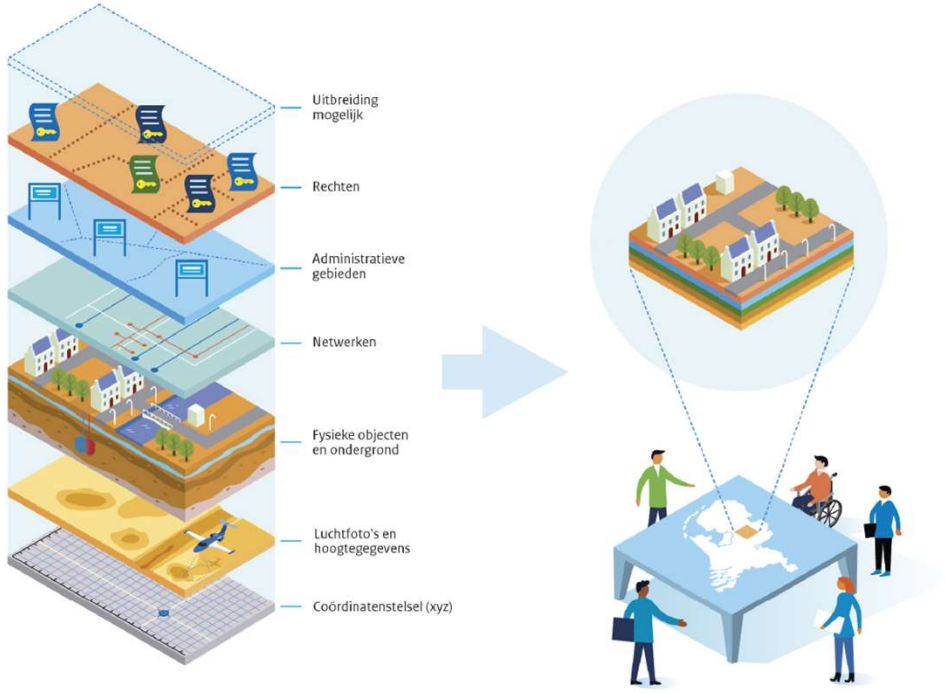
New registry with a coherent Geo view?

Emerging strategy on Federative data:

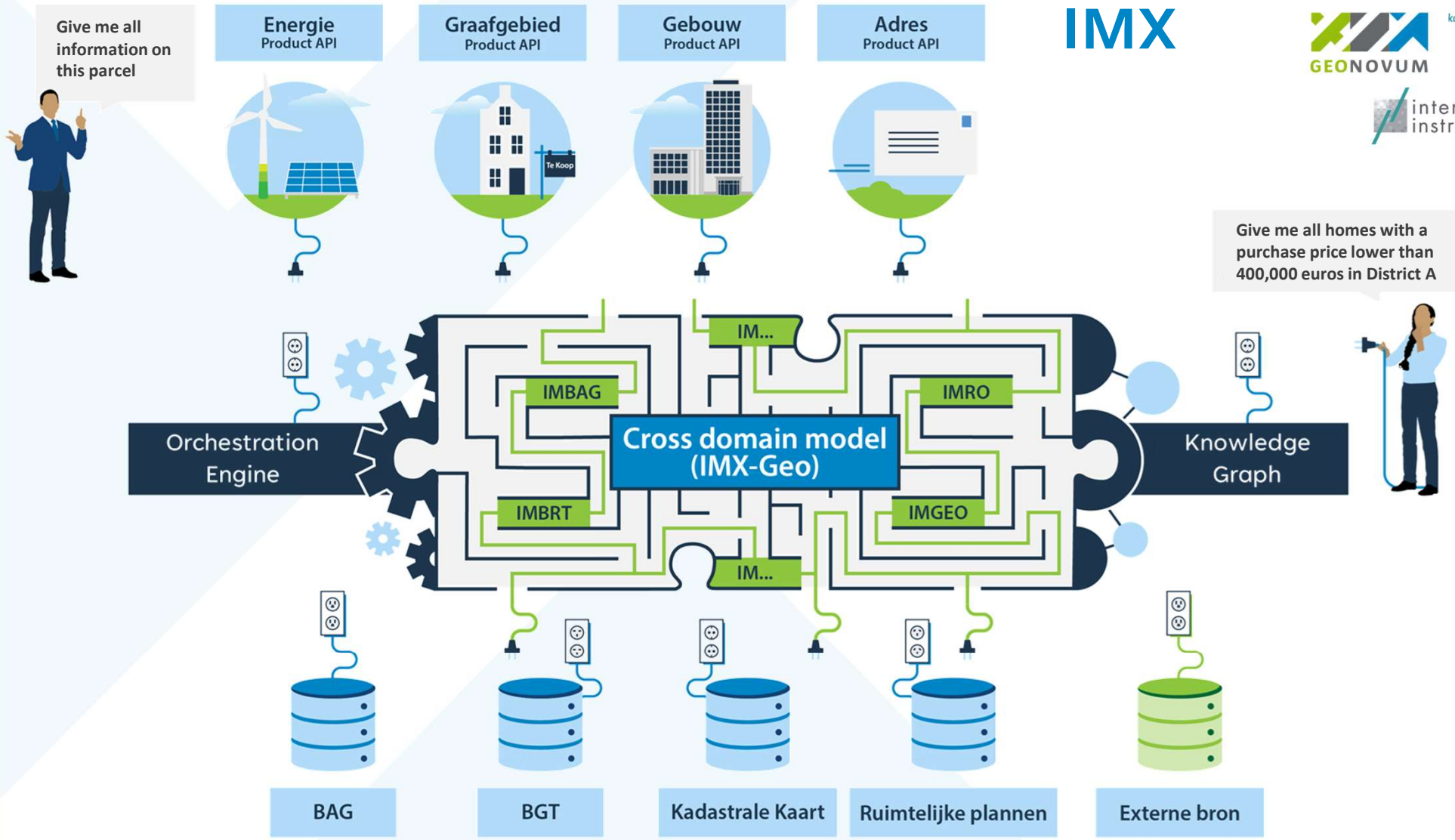
- data at the source
- stop copying data

Related programs:

- NGII
- Federatief datastelsel (IBDS)
- Zicht op Nederland (ZoN)
- Common Ground

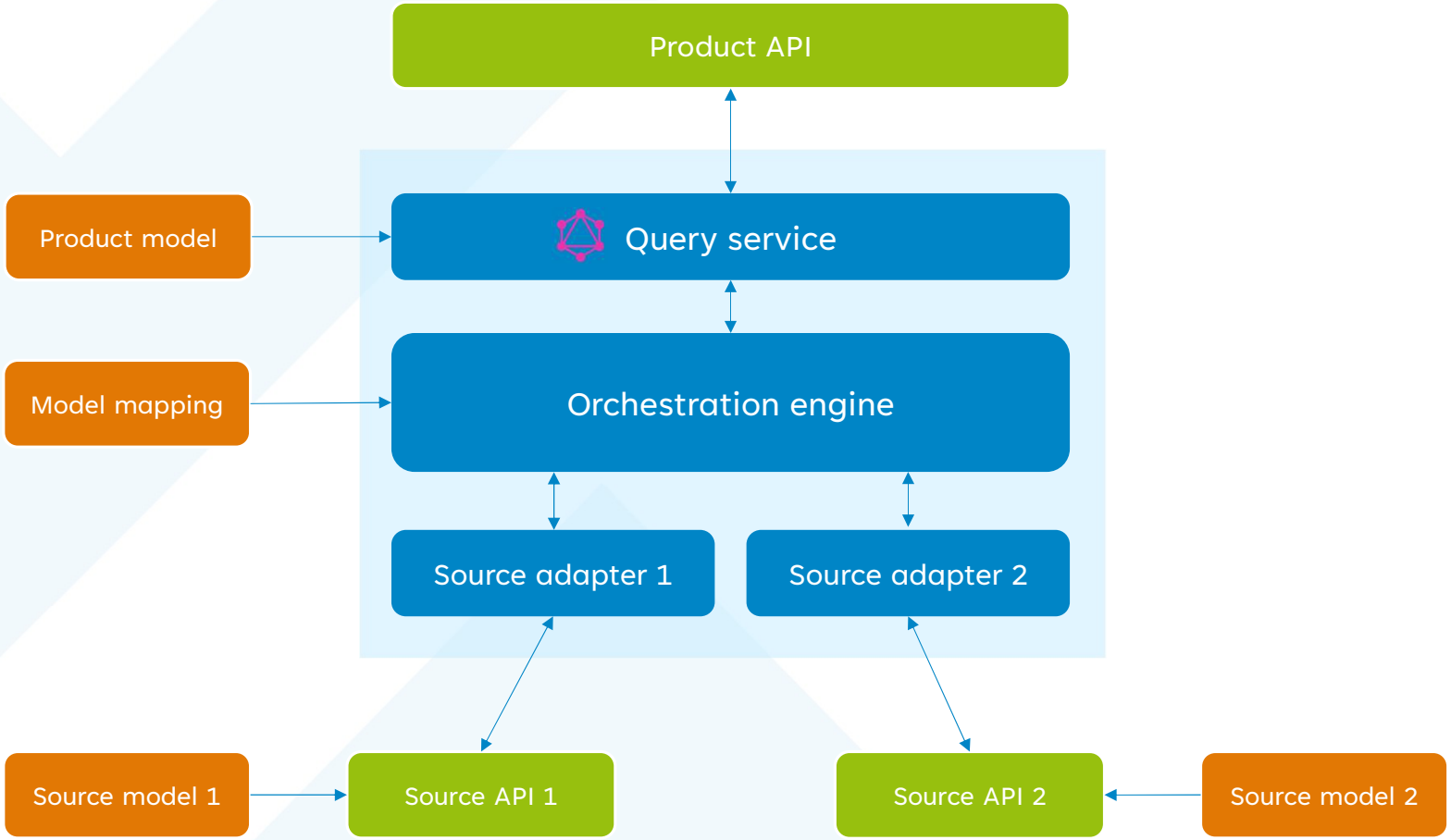


IMX



https://github.com/Geonovum/WaU

Orchestration engine



https://github.com/Geonovum/WaU

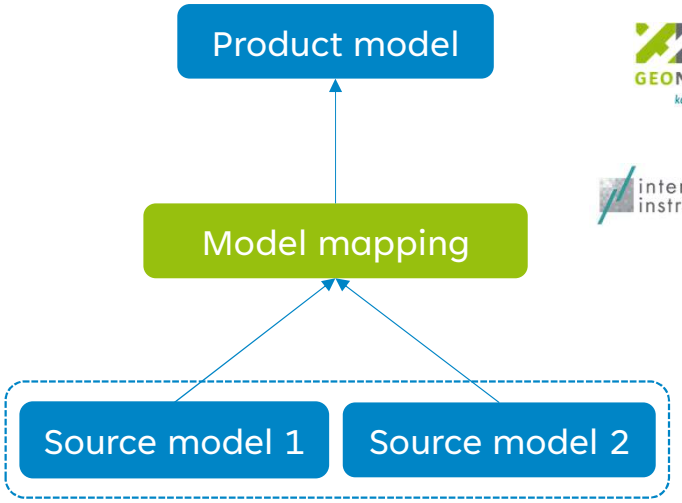
Product model: IMX-Geo

- Cross-domain model
- Combines the Geo base registries without changing that data sources
- Introduces new relationships between underlying concepts

https://github.com/Geonovum/WaU

IMX Model mapping

- Translates between product model and source model(s)
- Supports different modeling standards
  - MIM, OGC/ISO, RDFS/OWL/SHACL, ...
- Mapping and orchestration is stackable
  - A product model can become a source model
- Formal specification (emerging)
  - https://github.com/Geonovum/IMX-ModelMapping/




```
objectTypeMappings:
  Adres:
    sourceRoot: bag:Nummeraanduiding
    propertyMappings:
      identificatie:
        pathMappings:
          path: identificatie
      huisnummer:
        pathMappings:
          path: huisnummer
      huisnummertoevoeging:
        pathMappings:
          path: huisnummertoevoeging
      huisletter:
        pathMappings:
          path: huisletter
      postcode:
        pathMappings:
          path: postcode
```

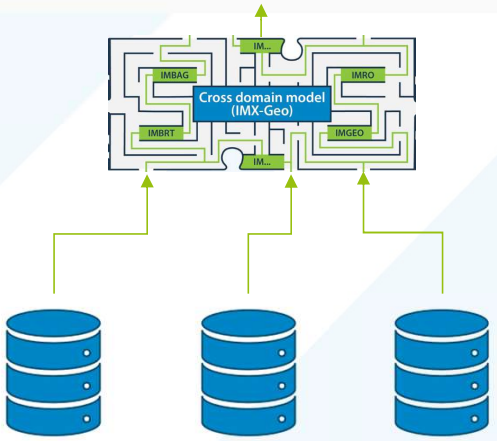
https://github.com/Geonovum/WaU

Orchestration Demo

Data Lineage

Data lineage of data objects

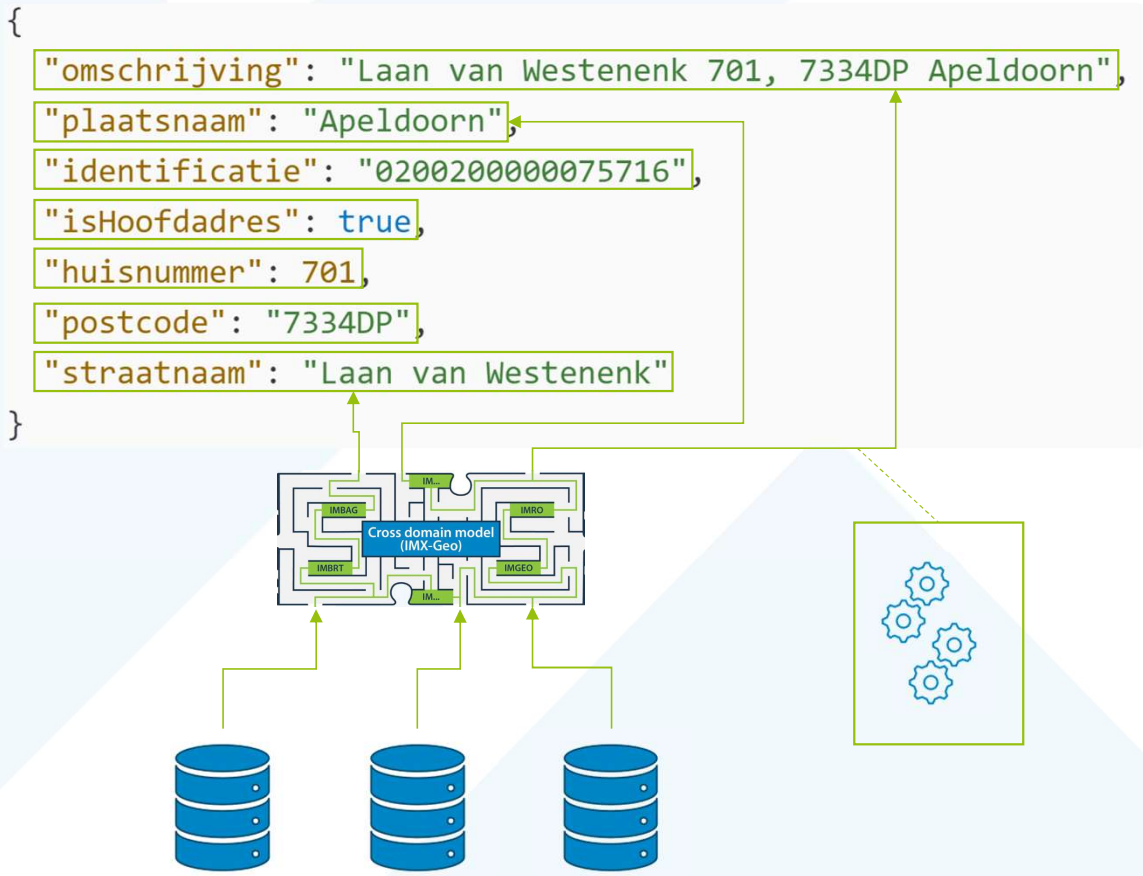

```
{
  "omschrijving": "Laan van Westenenk 701, 7334DP Apeldoorn",
  "plaatsnaam": "Apeldoorn",
  "identificatie": "0200200000075716",
  "isHoofdadres": true,
  "huisnummer": 701,
  "postcode": "7334DP",
  "straatnaam": "Laan van Westenenk"
}
```

- Representation of the path along which data flows from origin to usage.
- Describes data transformations and processing.
- Represented by interlinked components such as data elements, business processes, and IT systems.

https://github.com/Geonovum/WaU

Data lineage of data objects



- Representation of the path along which data flows from origin to usage.
- Describes data transformations and processing.
- Represented by interlinked components such as data elements, business processes, and IT systems.

https://github.com/Geonovum/WaU

Requirements for lineage of data elements

Data lineage is metadata

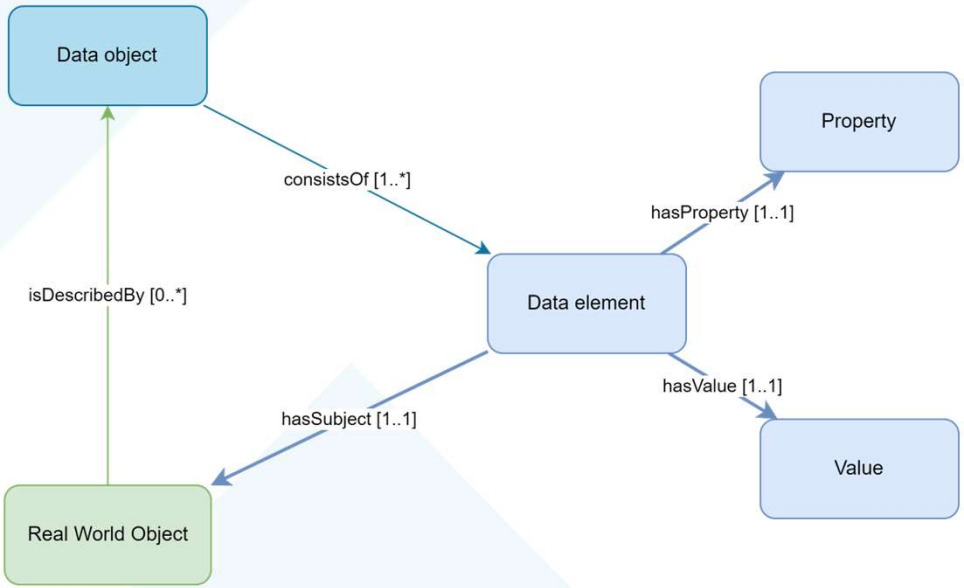
- We need a way to model data elements
- Yet, keep our core data models simple
- We want to keep lineage separate from the data object itself
- We want data lineage to be publishable separately from the data object

For this we need:

1. A way to model data elements as objects
2. A data model for data lineage

https://github.com/Geonovum/WaU

Modeling data elements as objects



https://github.com/Geonovum/WaU

Example: data object as data elements

```
{
  "identificatie": "0200200000075716",
  "plaatsnaam": "Apeldoorn",
  "isHoofdadres": true
}
```

```
[
  {
    "subject": {
      "identificatie": "0200200000075716"
    },
    "property": "identificatie",
    "value": {
      "stringValue": "0200200000075716"
    }
  },
  {
    "subject": {
      "identificatie": "0200200000075716"
    },
    "property": "plaatsnaam",
    "value": {
      "stringValue": "Apeldoorn"
    }
  },
  {
    "subject": {
      "identificatie": "0200200000075716"
    },
    "property": "isHoofdAdres",
    "value": {
      "booleanValue": true
    }
  }
]
```

https://github.com/Geonovum/WaU

Existing data lineage approaches

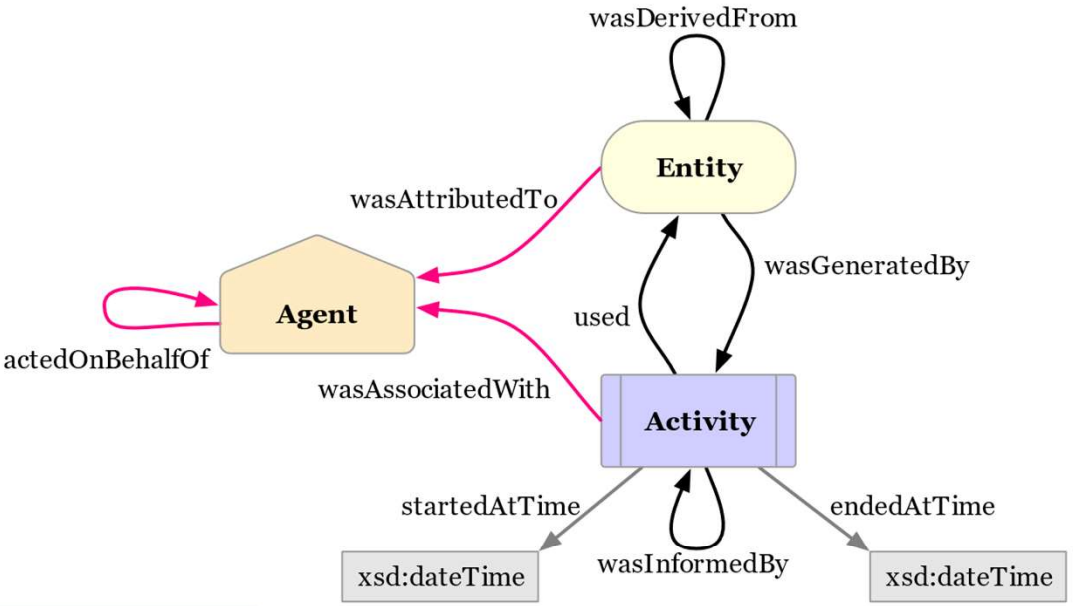
- Most data lineage approaches are tailored to data engineering in the data warehouse / data lake / data mesh space
- Job / run centric, focusing on columnar data

https://github.com/Geonovum/WaU

Data lineage based on W3C PROV

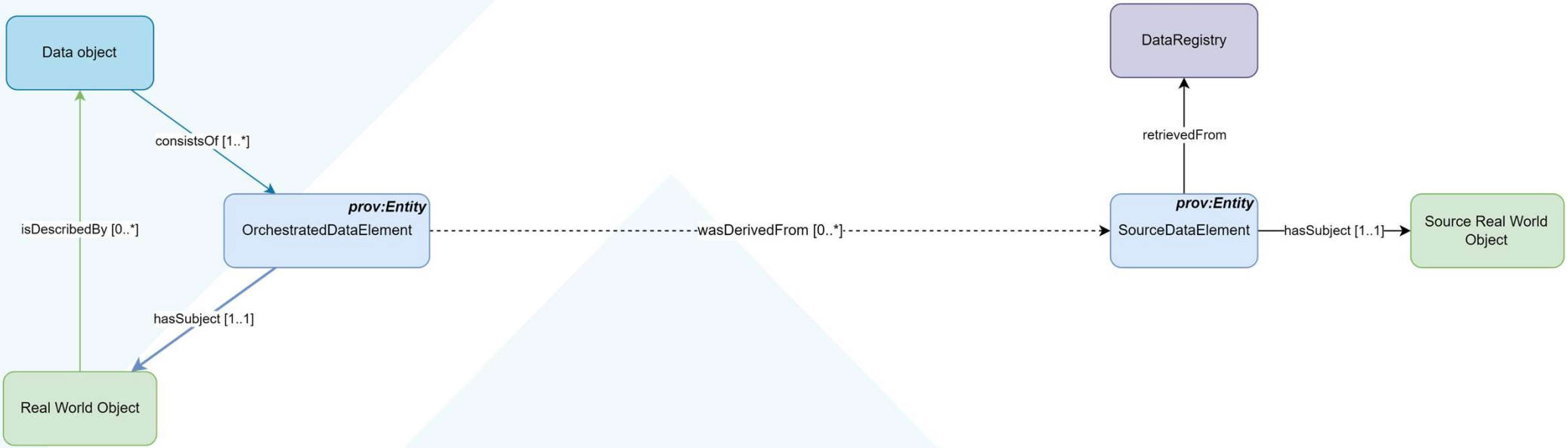
- Simple extensible model
- No presumptions on architecture

- Must specialize to use case



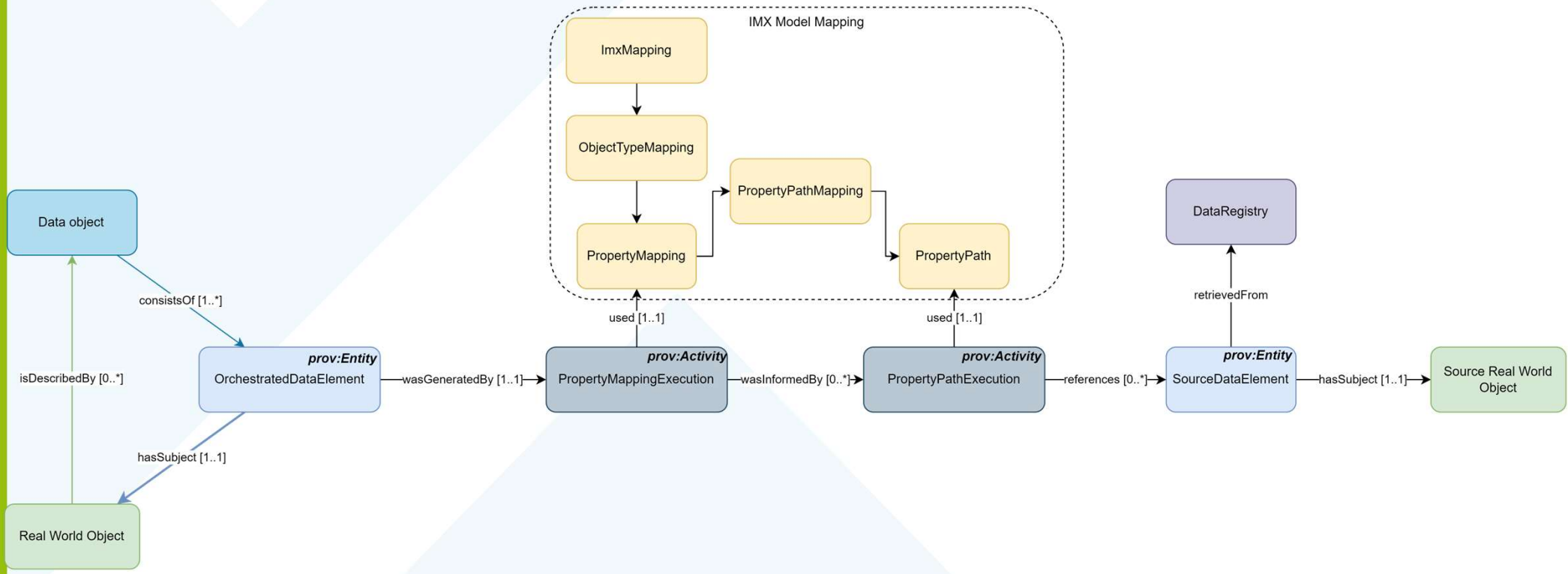
https://github.com/Geonovum/WaU

Data lineage based on W3C PROV



https://github.com/Geonovum/WaU

Data lineage based on W3C PROV



https://github.com/Geonovum/WaU

https://geonovum.github.io/IMX-Metadata/

https://geonovum.github.io/IMX-LineageModel

Publications

https://github.com/Geonovum/WaU

Data Lineage Demo

Future work

- Geonovum will organize an open consultation for IMX-Geo.
- IMX concept to be tested and further developed in https://digilab.overheid.nl/
- Talks with government agencies to develop the orchestration engine further. The orchestration engine will be open source.

  Follow development: https://github.com/imx-org/imx-orchestrate

Thanks!

Pano Maria

Contact

E-mail: p.maria@geonovum.nl
pano.maria@kadaster.nl
pano@skemu.com

www.geonovum.nl

https://github.com/Geonovum/WaU