golden agents

The Lenticular Lens

Addressing Various Aspects of Entity Disambiguation in the Semantic Web

Al Idrissou , Leon van Wissen , Veruska Zamborlini

Graphen und Netzwerke, 3 February 2022

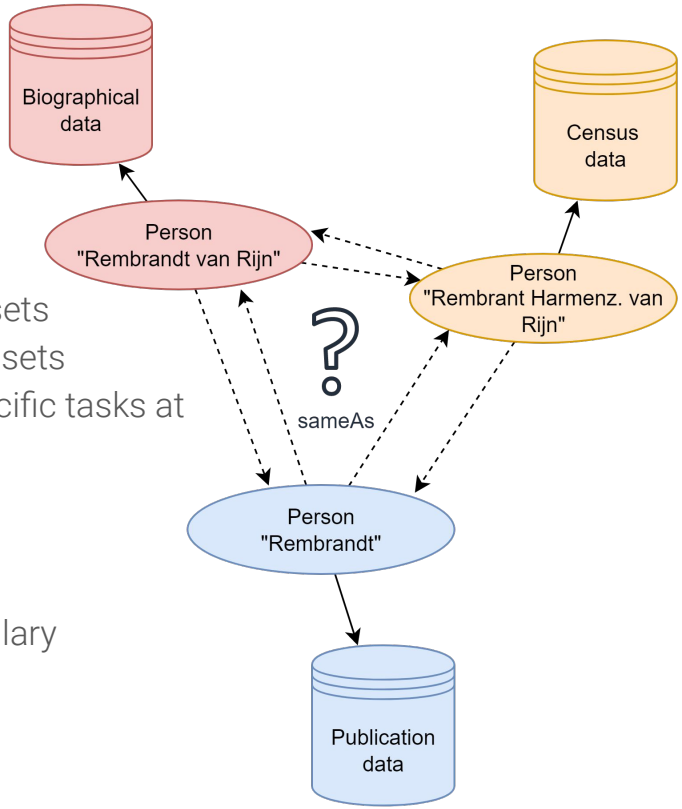
Objective

Problem

- Research data are **dispersed** over multiple (single-scoped) datasets
- Resources inside these data are usually **not linked** between datasets
- Existing tools are **too specific** or **not flexible enough** for the specific tasks at hand in the Golden Agents project

Solution

- **Generic yet flexible tool** that works with **RDF data** in *any* vocabulary
- Tailored 'rule-based **entity linking**'
- Off-the-shelf + Tailored + Ad-hoc **matching algorithms**
- **RDF Provenance** of matched links



The Lenticular Lens

Environment for entity disambiguation: Web interface that guides you in constructing **linksets** and **lenses**

- **Conditional Entity Selection** (i.e. combining selection criteria with AND/OR operators)
  - e.g. Persons in the role of author, or Books without an external identifier
- **Matching Methods**
  - Matching methods such as Levenshtein, Soundex, Jaro Winkler, Time delta, etc. + custom methods
  - They are implemented in **PostgreSQL** (PL/pgSQL)
  - Matching Results can be combined using **fuzzy logic** or **set-like operators**

https://github.com/knaw-huc/lenticular-lens

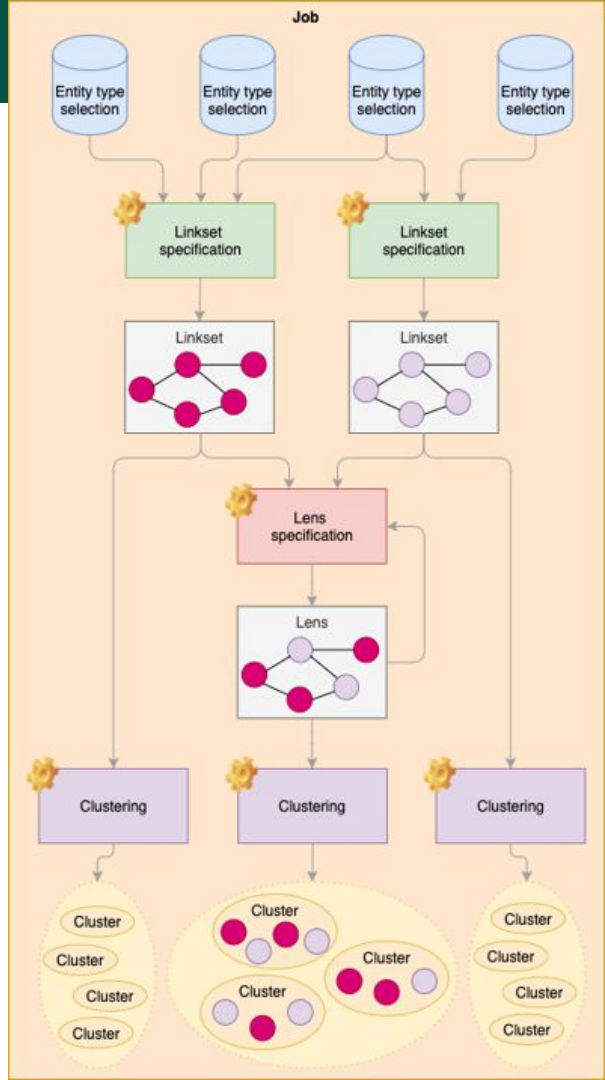
The Lenticular Lens

- **Clustering Methods**
  - Grouping linked entities into similarity clusters
- **Easy Validation**
  - Individual validation (manual)
  - Bulk validation
- **Exporting** capabilities: RDF in, RDF out (incl. provenance) or CSV
  - Extension to the VoID ontology: VOID+ (https://lenticularlens.org/docs/03.Ontology/#4-void-documentation)
  - Reification of the created links (as rdf:Statement, RDF*, or singletons)

https://github.com/knaw-huc/lenticular-lens

Overview

- **Creation :** Entity selection and linkset construction
- **Manipulation** : Combining linksets into lenses
- **Validation :** Per-link, cluster (manual or bulk)
- **Documentation :** Exporting links with provenance for reproducibility



Case Study: Golden Agents

Getting insight in the social or professional networks involved in the *production* and *consumption* of 'Occasional Poetry'.

Datasets

Golden Agents (origin: KB The Hague)

- Occasional Poetry: Publications written for one or more Persons on the occasion of a particular Event
  e.g. "**Poem** on the **Marriage** of **P.C. Hooft** and **Heleonora Hellemans** on **1627-11-30**"

Golden Agents Archival Documents (origin: City Archives Amsterdam)

- Index on Notices of Marriage:
  e.g. "**Record** on the **Notice of Marriage** of **P.C. Hooft** (groom) and **Leonora Hellemans** (bride) on **1627-11-03**"
- Index on Baptisms:
  e.g. "**Record** on the **Baptism** of Christina (child) with **P.C. Hooft** (father) and **Eleonora Hellemans** (mother) on **1628-08-20**"

Setting up a match

Entity selection

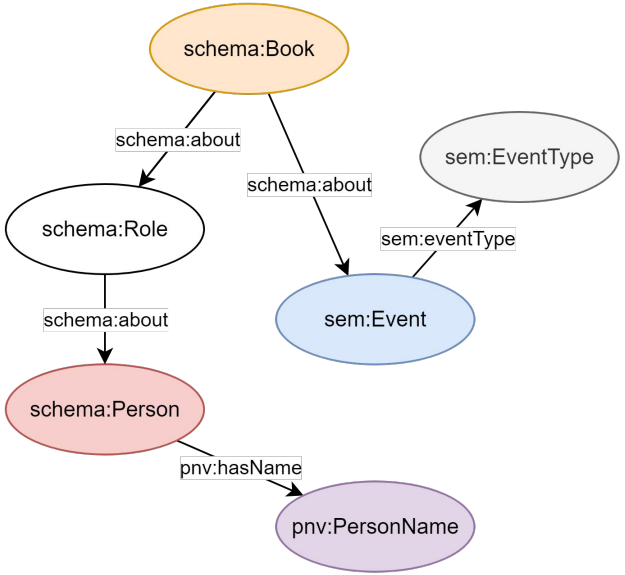
In RDF, this is what is expressed in the `rdf:type` property

- Occasional Poetry: Person (`schema:Person`)
- City Archives data: Person (`ga:Person`)

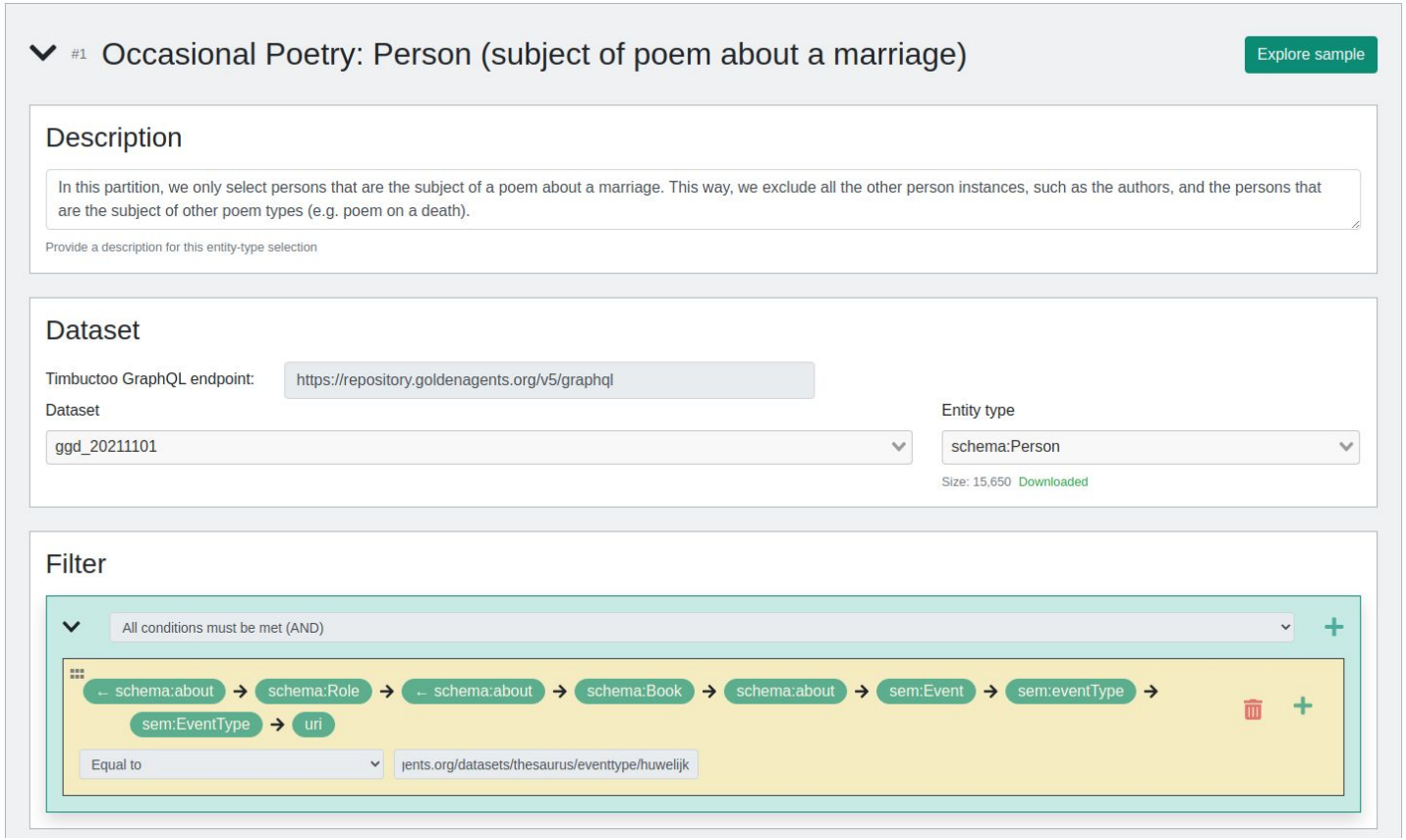
Filtering

In this example, only on persons in poems on Marriages

- 'Walking' the graph cf. property path



Vocabularies:
schema: http://schema.org/
sem: http://semanticweb.cs.vu.nl/2009/11/sem/
pnv: https://w3id.org/pnv#



Data Partition: Selecting which entity (selection, based on filters) is matched on/to. The filter is a property path.

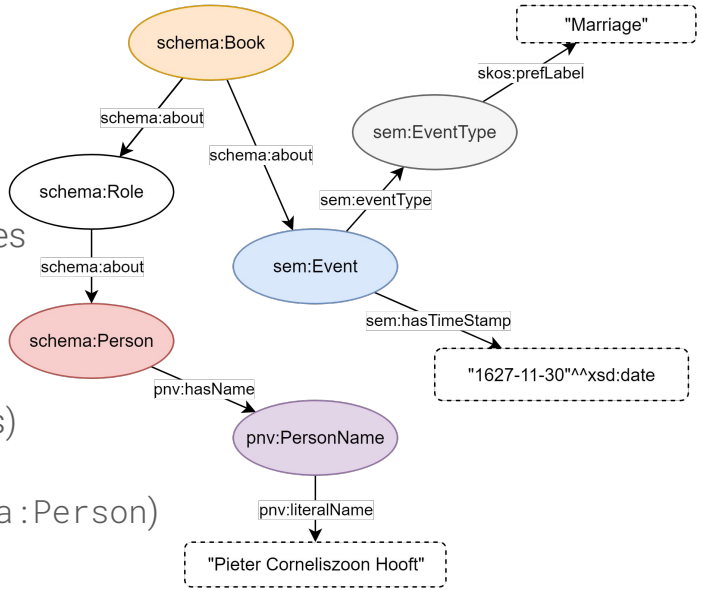
Setting up a match

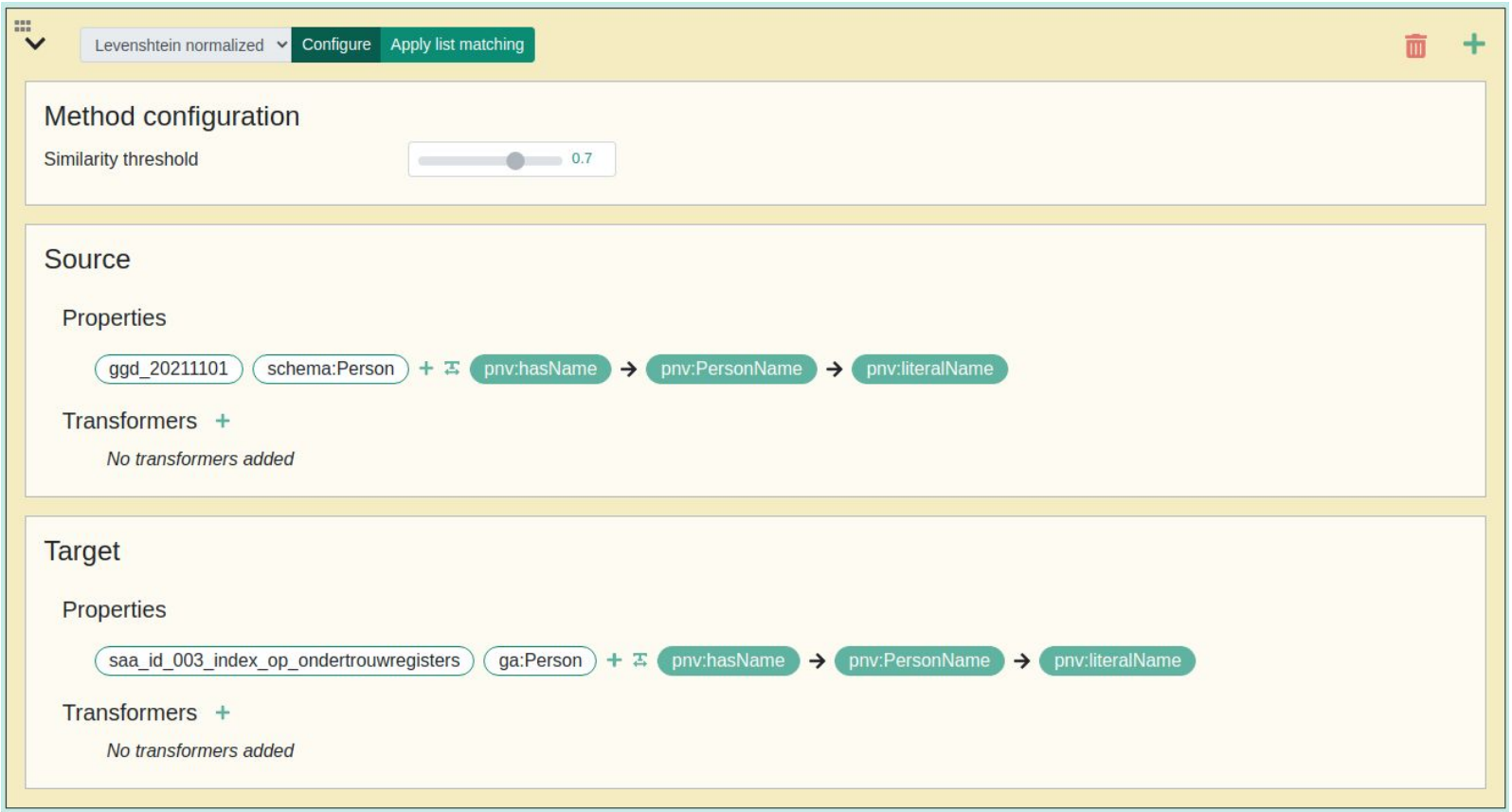
Specifying matching rules

Linking the Occasional Poetry data to the Notice of Marriage registries

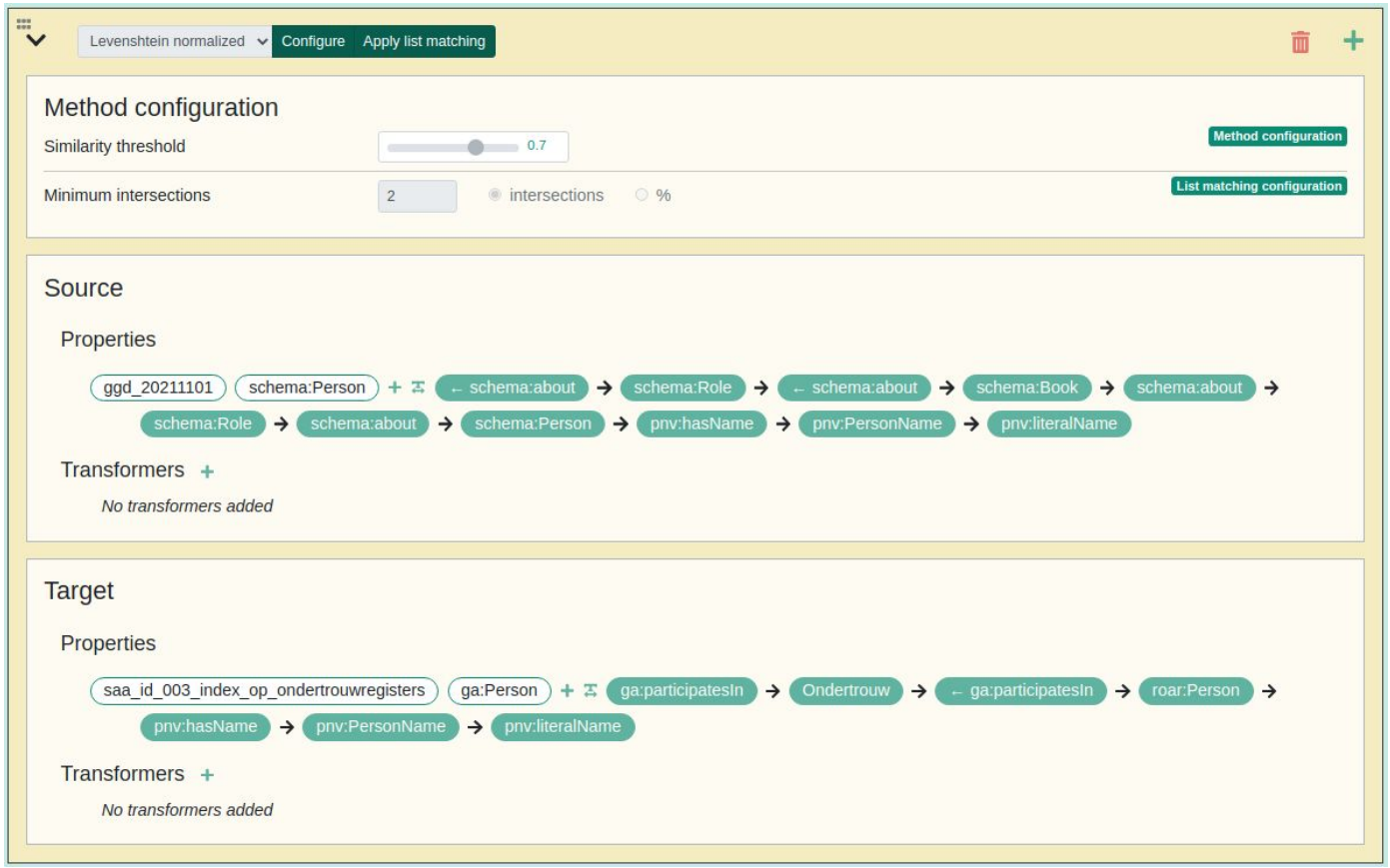
1. String: Name of the person is similar (fuzzy)
2. String: Name of the partner is similar (fuzzy)
3. Date: Notice of Marriage before Marriage (time delta 6 months)

Specifying this for the Source (`schema:Person`) and Target entity (`ga:Person`)

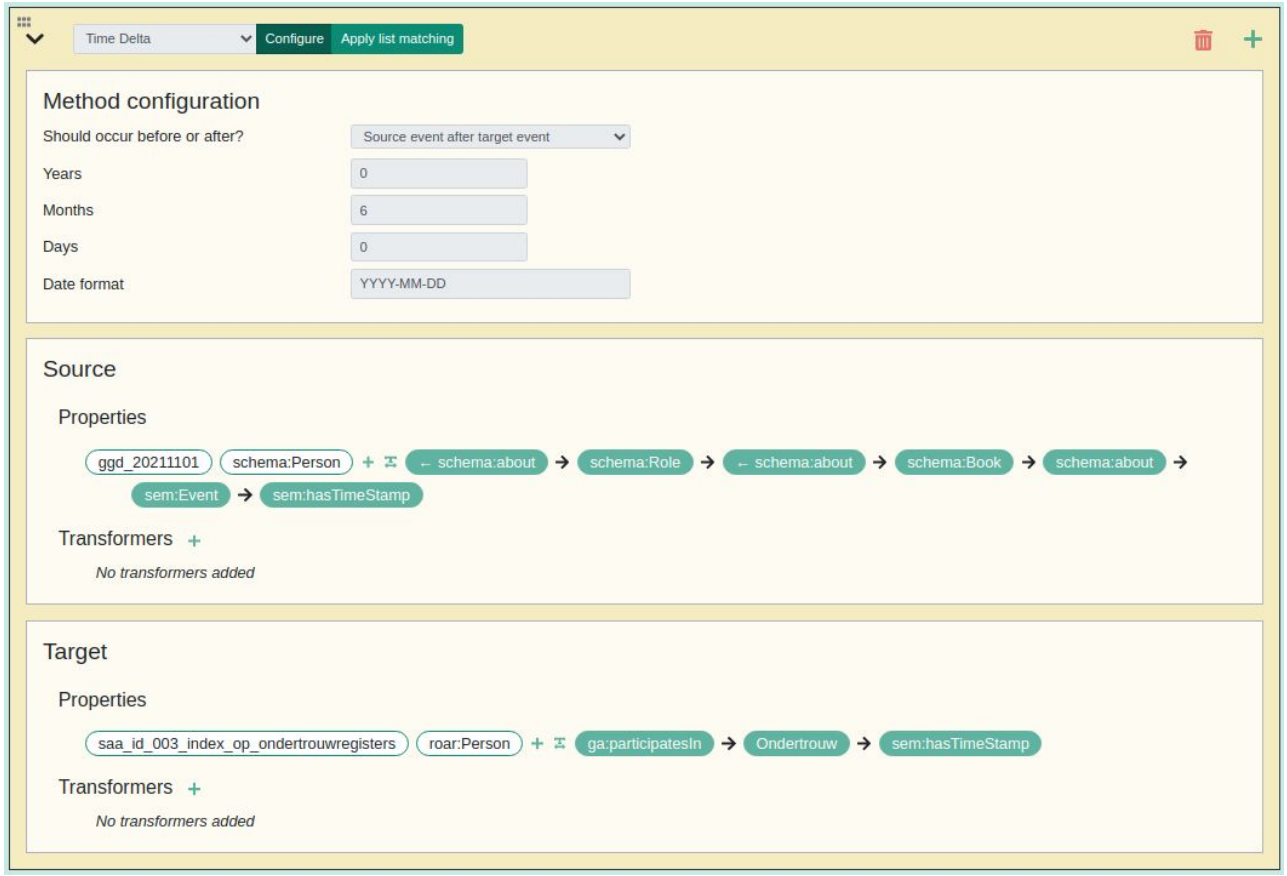




String matching: Levenshtein (normalized) 0.7 on the name of the person



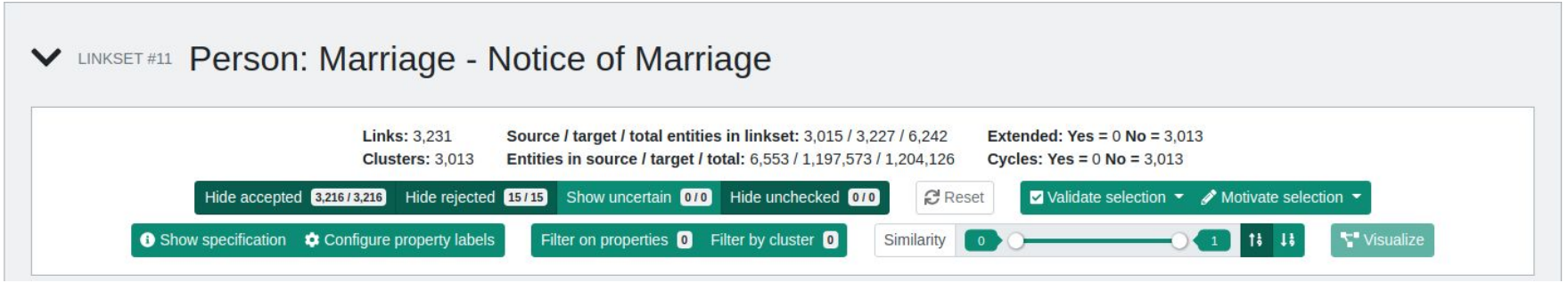
String matching: Levenshtein (normalized) 0.7 on the name of the person + the name of the partner. At least 2 of the names in the source should overlap with the names in the target.



Time delta: Only match against Persons in Events no later than 6 months apart.
The source (Marriage) occurs after the target (Notice of Marriage).

Result

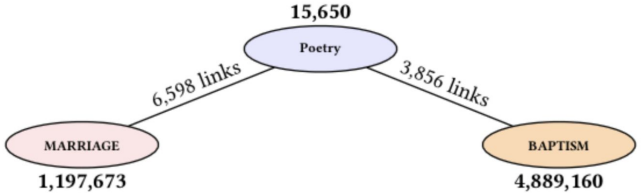
- Comparing 6,533 persons (source) to 1,197,573 persons (target) took ~7 hours
- 3,231 links were found
- 3,013 clusters were found

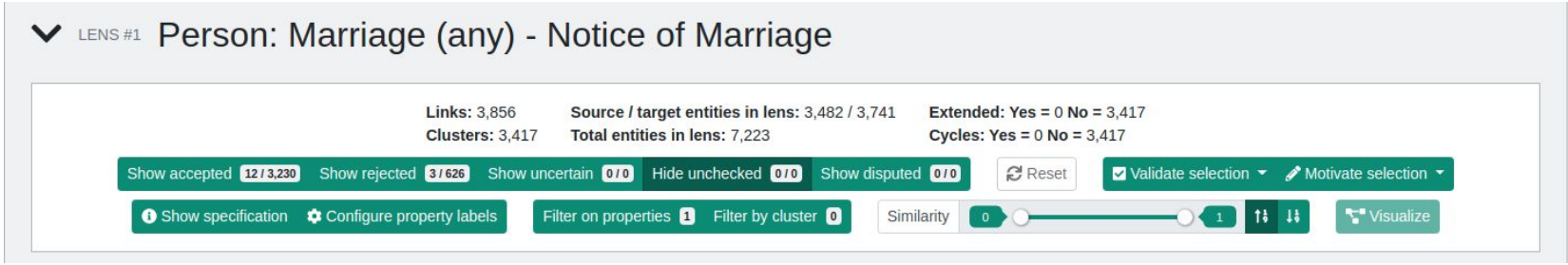


Lenses

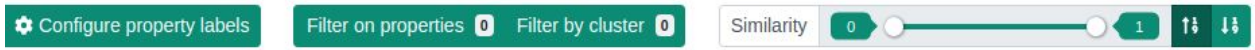
Combining multiple linksets

- Different matching strategies for different data:
  - 6 months difference for Marriages
  - 50 years difference for Marriage Anniversaries
  - 20 years for Baptisms
- Prevents duplicate validation
- Exporting in a single linkset



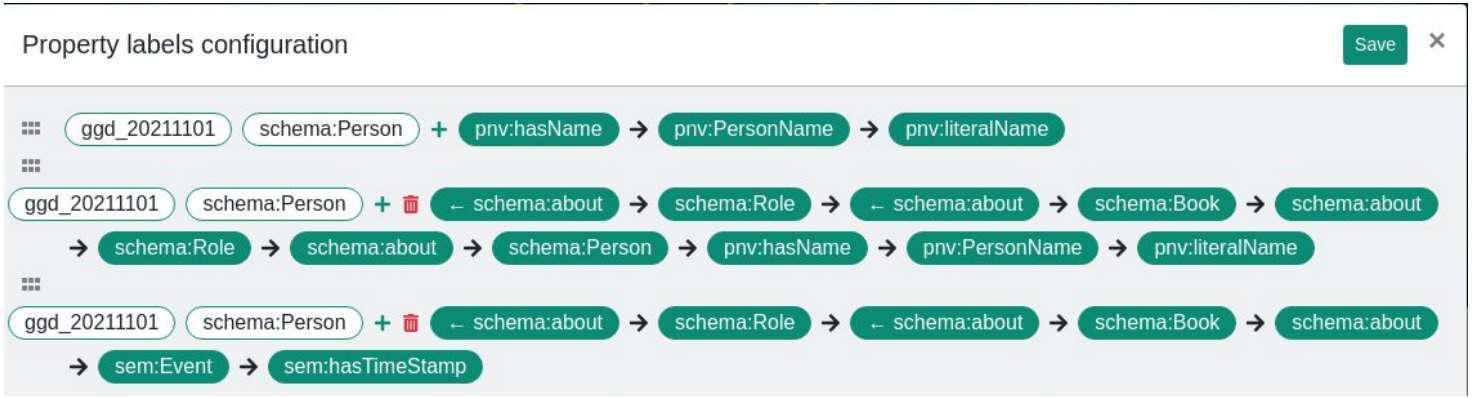


Validation

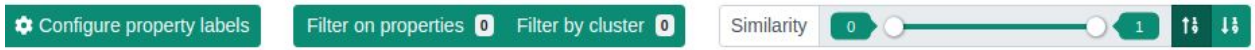


Extra information

- Specify what other information is useful for the (manual) validation process

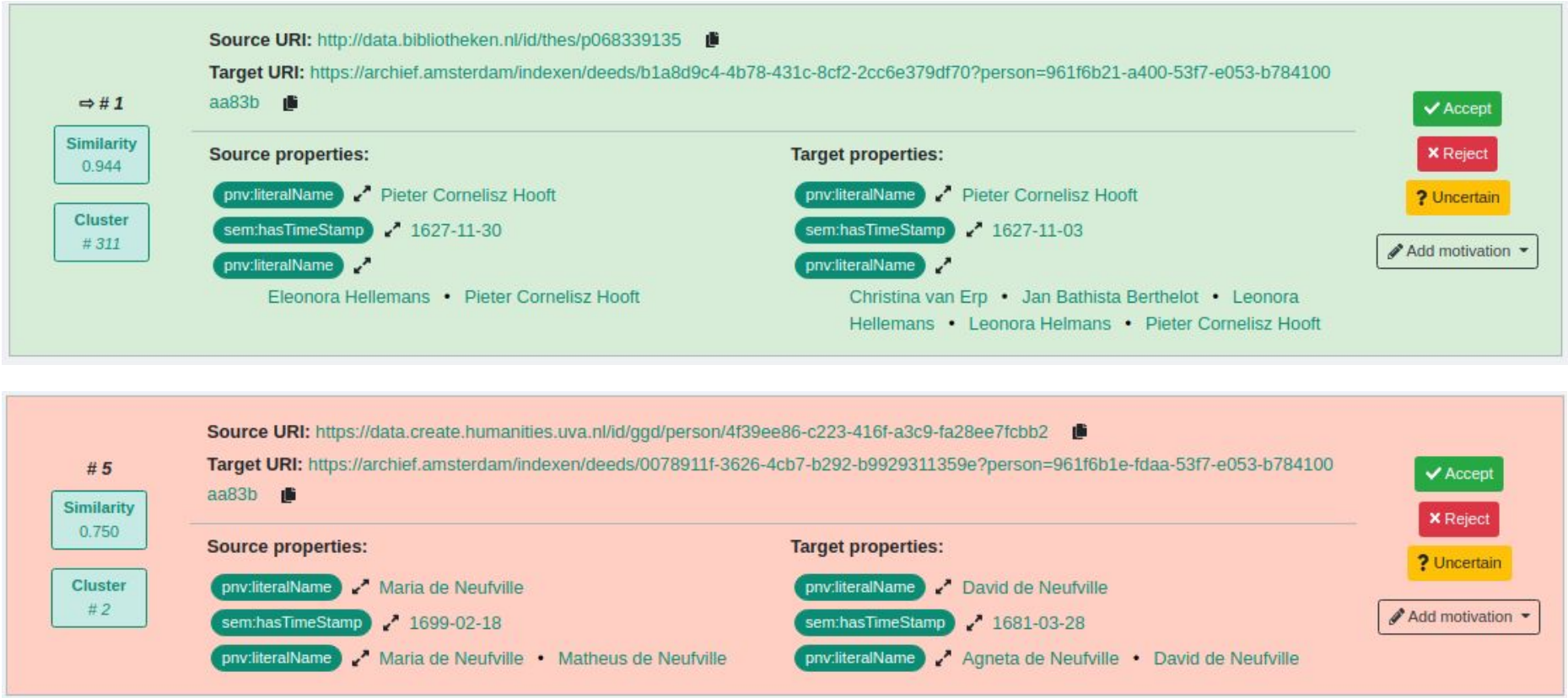


Validation



Filtering

- Similarity
  - Based on similarity score, e.g. accept all above 0.9
- Properties
  - Search tool, e.g. quickly find an entity by name
  - Error spotting: e.g. filtering out all links where a person in a groom role has been linked to a bride role
- Clusters
  - Validate per cluster, e.g. the entire family 'Ploos van Amstel' in one go



Accepted (above) and rejected (below) link

VoidPlus documentation: <https://lenticularlens.org/docs/03.Ontology/#4-void-documentation

A single link from this data

Single link

```
<http://data.bibliotheken.nl/id/thes/p068339135>    owl:sameAs
<https://archief.amsterdam/indexen/deeds/b1a8d9c4-4b78-431c-8cf2-2cc6e379df70?person=961f6b21-a400
-53f7-e053-b784100aa83b> .
```

Standard Reification (statement)

```
resource:Reification-4b7937d4d1d0b62
    a                                       rdf:Statement ;
    rdf:predicate                           owl:sameAs ;
    rdf:subject                             <http://data.bibliotheken.nl/id/thes/p068339135> ;
    rdf:object
<https://archief.amsterdam/indexen/deeds/b1a8d9c4-4b78-431c-8cf2-2cc6e379df70?person=961f6b21-a400
-53f7-e053-b784100aa83b> ;
    voidPlus:matchingStrength               "0.94444"^^xsd:decimal ;
    voidPlus:hasClusterID                   cluster:49686f5b907d8f3 ;
    voidPlus:hasValidation                  validation:4b7937d4d1d0b62 .
```

VoidPlus documentation: <https://lenticularlens.org/docs/03.Ontology/#4-void-documentation

A single link from this data

Reification (cluster)

```
cluster:49686f5b907d8f3
    a                                   voidPlus:Cluster ;
    voidPlus:nodes                      "2"^^xsd:integer ;
    voidPlus:hasItem                    <http://data.bibliotheken.nl/id/thes/p068339135> ,
<https://archief.amsterdam/indexen/deeds/b1a8d9c4-4b78-431c-8cf2-2cc6e379df70?person=961f6b21-a400
-53f7-e053-b784100aa83b> .
```

Reification (validation)

```
validation:4b7937d4d1d0b62
     a                                   voidPlus:Validation ;
     voidPlus:hasValidationStatus        resource:Accepted .
```

Other use cases in the project

Persons

- Connecting occurrences in Baptism, Marriage, Burial registries to references of persons in the Notarial Archives of Amsterdam

Books

- Linking individual book entries in a probate inventory to books in the STCN

Paintings

- Linking works in the Rijksmuseum to a collection of Group Portraits (and to who's depicted)

Contact

Al Idrissou
alkoudouss@yahoo.com

Leon van Wissen
l.vanwissen@uva.nl

Veruska Zamborlini
veruska.zamborlini@ufes.br

Acknowledgements
https://www.goldenagents.org/staff/

Golden Agents
https://www.goldenagents.org/

Clariah
https://www.clariah.nl/

RISIS
http://risis.eu/

https://github.com/knaw-huc/lenticular-lens