

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

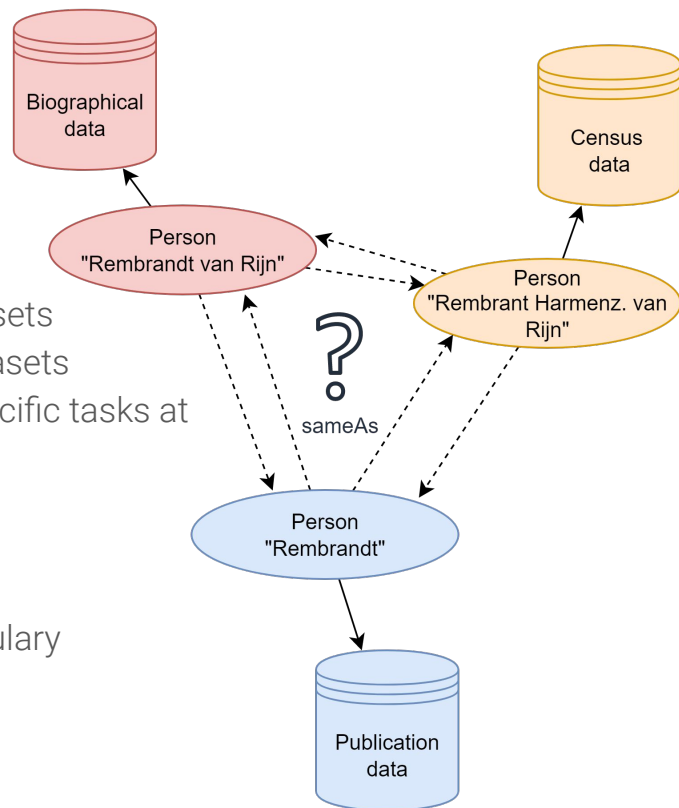
Objective

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**

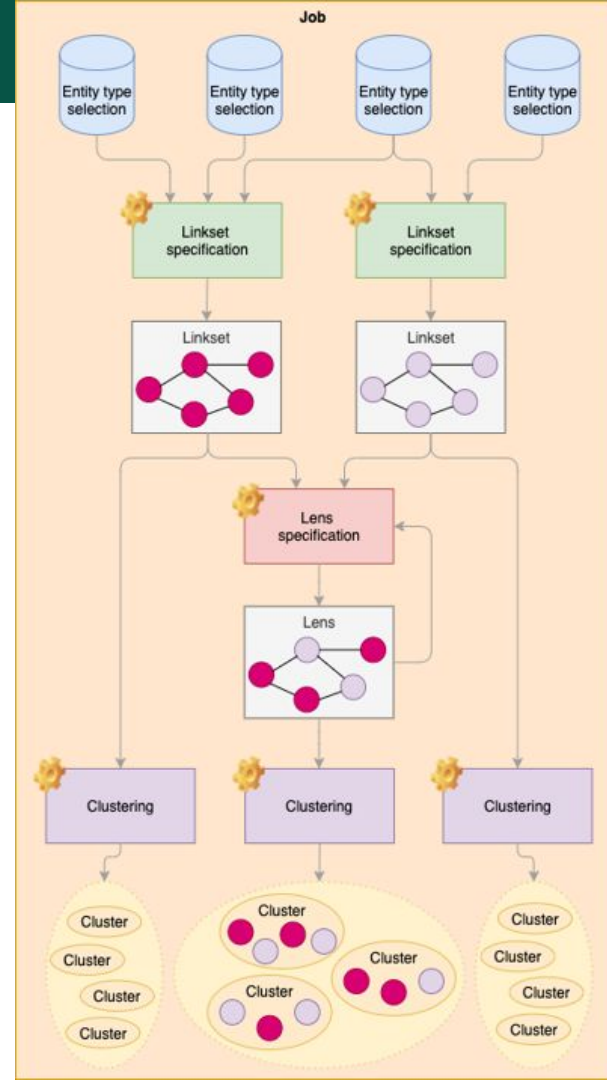


The Lenticular Lens

- **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)

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

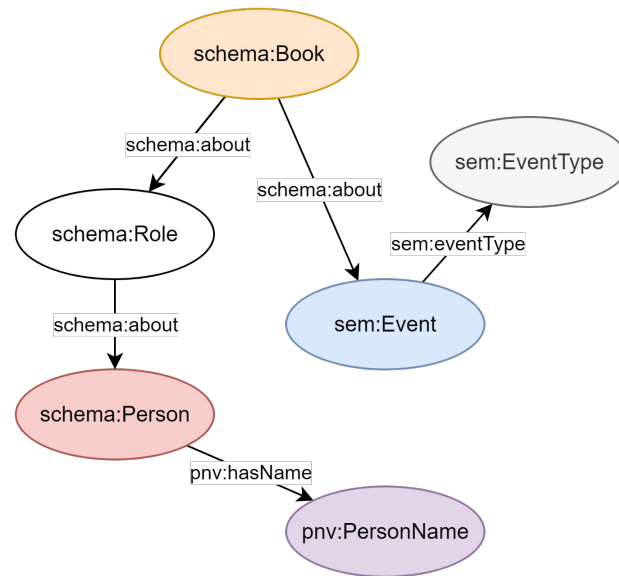
In RDF, this is what is expressed in the `rdf:type` property

- 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#>

▼ #1 Occasional Poetry: Person (subject of poem about a marriage)

[Explore sample](#)

Description

In this partition, we only select persons that are the subject of a poem about a marriage. This way, we exclude all the other person instances, such as the authors, and the persons that are the subject of other poem types (e.g. poem on a death).

Provide a description for this entity-type selection

Dataset

Timbuctoo GraphQL endpoint:

<https://repository.goldenagents.org/v5/graphql>

Dataset

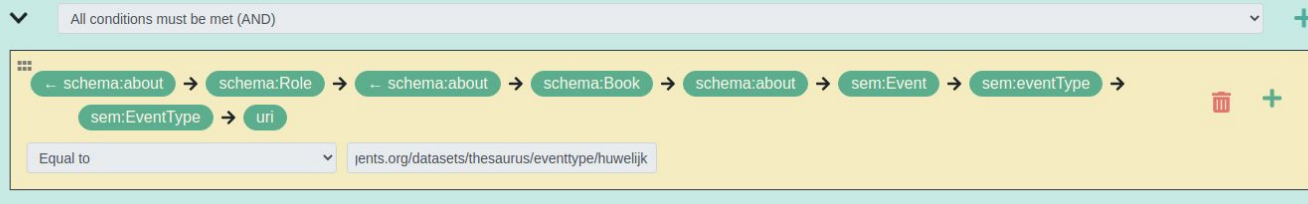
ggd_20211101

Entity type

schema:Person

Size: 15,650 [Downloaded](#)

Filter



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

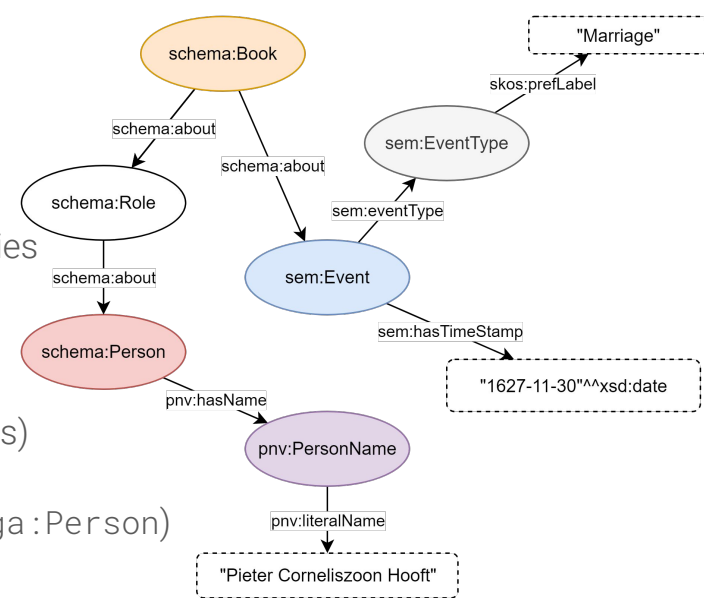
Setting up a match

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)



Levenshtein normalized

Configure

Apply list matching

Method configuration

Similarity threshold

Source

Properties

ggd_20211101 schema:Person + ↔ pnv:hasName → pnv:PersonName → pnv:literalName

Transformers +

No transformers added

Target

Properties

saa_id_003_index_op_ondertrouwregisters ga:Person + ↔ pnv:hasName → pnv:PersonName → pnv:literalName

Transformers +

No transformers added

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

Levenshtein normalized

Configure

Apply list matching

Method configuration

Similarity threshold

0.7

Method configuration

Minimum intersections

2

☒ intersections

☐ %

List matching configuration

Source

Properties

ggd_20211101

schema:Person

+

↕

← schema:about

→

schema:Role

→

← schema:about

→

schema:Book

→

schema:about

→

schema:Role

→

schema:about

→

schema:Person

→

pnv:hasName

→

pnv:PersonName

→

pnv:literalName

Transformers +

No transformers added

Target

Properties

saa_id_003_index_op_ondertrouwregisters

ga:Person

+

↕

ga:participatesIn

→

Ondertrouw

→

← ga:participatesIn

→

roar:Person

→

pnv:hasName

→

pnv:PersonName

→

pnv:literalName

Transformers +

No transformers added

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

Configure

Apply list matching

Method configuration

Should occur before or after?

Source event after target event

Years

0

Months

6

Days

0

Date format

YYYY-MM-DD

Source

Properties

ggd_20211101

schema:Person

 +

← schema:about

 →

schema:Role

 →

← schema:about

 →

schema:Book

 →

schema:about

 →

sem:Event

 →

sem:hasTimeStamp

Transformers +

No transformers added

Target

Properties

saa_id_003_index_op_ondertrouwregisters

roar:Person

 +

ga:participatesIn

 →

Ondertrouw

 →

sem:hasTimeStamp

Transformers +

No transformers added

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

LINKSET #11 Person: Marriage - Notice of Marriage

Links: 3,231

Clusters: 3,013

Source / target / total entities in linkset: 3,015 / 3,227 / 6,242

Entities in source / target / total: 6,553 / 1,197,573 / 1,204,126

Extended: Yes = 0 No = 3,013

Cycles: Yes = 0 No = 3,013

Hide accepted 3,216 / 3,216

Hide rejected 15 / 15

Show uncertain 0 / 0

Hide unchecked 0 / 0

Reset

☒ Validate selection

Motivate selection

Show specification

Configure property labels

Filter on properties 0

Filter by cluster 0

Similarity

0

1

↑↓

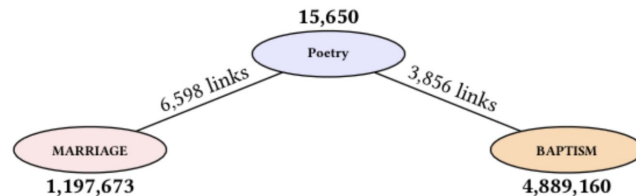
↑↓

Visualize

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



▼ LENS #1 Person: Marriage (any) - Notice of Marriage

Links: 3,856
Clusters: 3,417

Source / target entities in lens: 3,482 / 3,741
Total entities in lens: 7,223

Extended: Yes = 0 No = 3,417
Cycles: Yes = 0 No = 3,417

Show accepted 12 / 3,230

Show rejected 3 / 626

Show uncertain 0 / 0

Hide unchecked 0 / 0

Show disputed 0 / 0

Reset

☒ Validate selection

Motivate selection

Show specification

Configure property labels

Filter on properties 1

Filter by cluster 0

Similarity

0 1

↑ ↓

Visualize

Validation

⚙️ Configure property labels

Filter on properties 0

Filter by cluster 0

Similarity

0



1



Extra information

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

Property labels configuration

Save



ggd_20211101 schema:Person + pnv:hasName → pnv:PersonName → pnv:literalName

ggd_20211101 schema:Person + schema:about → schema:Role → schema:about → schema:Book → schema:about
→ schema:Role → schema:about → schema:Person → pnv:hasName → pnv:PersonName → pnv:literalName

ggd_20211101 schema:Person + schema:about → schema:Role → schema:about → schema:Book → schema:about
→ sem:Event → sem:hasTimeStamp



Validation

⚙️ Configure property labels

Filter on properties 0

Filter by cluster 0

Similarity

0  1 

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

Source URI: <http://data.bibliotheken.nl/id/thes/p068339135> 

Target URI: <https://archieff.amsterdam/indexen/deeds/b1a8d9c4-4b78-431c-8cf2-2cc6e379df70?person=961f6b21-a400-53f7-e053-b784100aa83b> 

⇒ # 1

Similarity
0.944

Cluster
311

Source properties:

pnv:literalName ↗ Pieter Cornelisz Hooft

sem:hasTimeStamp ↗ 1627-11-30

pnv:literalName ↗

Eleonora Hellemans • Pieter Cornelisz Hooft

Target properties:

pnv:literalName ↗ Pieter Cornelisz Hooft

sem:hasTimeStamp ↗ 1627-11-03


pnv:literalName ↗


Christina van Erp • Jan Bathista Berthelot • Leonora Hellemans • Leonora Helmans • Pieter Cornelisz Hooft


✓ Accept

✗ Reject

? Uncertain

 Add motivation ▼

Source URI: <https://data.create.humanities.uva.nl/id/ggd/person/4f39ee86-c223-416f-a3c9-fa28ee7fcbb2> 

Target URI: <https://archieff.amsterdam/indexen/deeds/0078911f-3626-4cb7-b292-b9929311359e?person=961f6b1e-fdaa-53f7-e053-b784100aa83b> 

5

Similarity
0.750

Cluster
2

Source properties:

pnv:literalName ↗ Maria de Neufville

sem:hasTimeStamp ↗ 1699-02-18

pnv:literalName ↗ Maria de Neufville • Matheus de Neufville

Target properties:

pnv:literalName ↗ David de Neufville


sem:hasTimeStamp ↗ 1681-03-28

pnv:literalName ↗ Agneta de Neufville • David de Neufville

✓ Accept

✗ Reject

? Uncertain

 Add motivation ▼

Accepted (above) and rejected (below) link

A single link from this data

Single link

```
<http://data.bibliotheken.nl/id/thes/p068339135>      owl:sameAs
<https://archieff.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://archieff.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 .
```



A single link from this data

Reification (cluster)

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

Reification (validation)

```
validation:4b7937d4d1d0b62
  a
  voidPlus:hasValidationStatus
    voidPlus:Validation ;
    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/>



RISIS
Research infrastructure for research
and innovation policy studies



Clariah



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