The Lenticular Lens

Addressing Various Aspects of Entity Disambiguation in the Semantic Web

Al Idrissou[1], Leon van Wissen[2], and Veruska Zamborlini[3]

[1]Vrije Universiteit Amsterdam, The Netherlands
[2]University of Amsterdam, The Netherlands
[3]Federal University of Espirito Santo, Brazil

Figure 1: An environment for aligning resources across datasets and explicit decoration of the resulting links and sets of links.

1 Introduction

Time and again, researchers are presented with problems for which they postulate and test hypotheses in order to provide us with robust explanations for research questions successfully investigated. Oftentimes, solid explanations for complex problems require exploring a multitude of data-sources rather than the use of a unique (authoritative) data-source. This is the case, for example, in domains such as e-(science/commerce), tourism, digital humanities, etc. In the Digital Humanities, in particular, answering a research question often requires a researcher to combine several datasets that each contain specific and single scoped information on the entities it describes/contains. However, the use of multiple sources comes with a problem of its own: data integration.

In the Semantic Web, entity matching is a well-known technique for which there exists a considerable body of work on generic matching algorithms (Hassanzadeh et al., 2009,?; Mirani and Radke, 2014; Ngomo and Auer, 2011; Volz et al., 2009a) dedicated for the task of integrating data. Only, these are scattered and solely generate links between a pair of datasets. This triggers the need for tailoring and/or combining algorithms which is not an unusual reality for resource linking.

Once executed, this computationally expensive task of contextual data integration should record the processes that lead to the discovery of the links. Only, developers of matching algorithms do not concern themselves with such metadata acquisition, leaving data providers to resort to time-consuming and error-prone manual labor for describing their data. When available, the metadata is often provided using VoID

(Alexander et al., 2009), the W3C standard (Alexander et al., 2011) vocabulary for interlinked datasets. However, this does not provide enough means to describe the full provenance of matching, combining, clustering, and validating links in manners that fully support assessing the data for reliability, reuse, reproducibility, or quality.

We present the Lenticular Lens, a user-friendly web interface tool that provides a set of means (data linking / integration) to an end: answering data-driven research questions (data analysis). It offers a context-dependent user-guided entity matching across multiple datasets using ad-hoc and/or off-the-shelf generic algorithms that can be logically combined. Such combination is performed over the scores of the links discovered by the various user-selected algorithms using a set of provided fuzzy logic operators. In the end, it utilizes our proposed VoID+ ontology[1], to intelligibly document all user-defined processes leading to the discovery, manipulation, clustering, and validation of links. All these processes can later on, at the user's convenience, be exported in various formats that include RDF and CSV.

2 The Lenticular Lens

A first version of the Lenticular Lens tool (Idrissou et al., 2018) was developed by Al Idrissou at the Vrije Universiteit Amsterdam and was further developed as part of the Golden Agents project[2] in which the tool is used to interconnect resources from various heterogeneous datasets on cultural production and consumption in 17^{th} and 18^{th} century Amsterdam. With it, users can decide on what, how and when to link; they can cluster the matched resources, manipulate and/or validate the discover links; they can export the links with or without their respective metadata.

3 Hands-on

To have a feel of the Lenticular Lens, this section introduces a use case through which we elaborate on the main features of the tool. These features enforce the explicit specification and documentation of processes that span from data partitioning to link creation (linkset), manipulation (linkset) and validation. To help smoothing out validation, the tool enables users to rely on techniques such as clustering and/or visualization.

3.1 Golden Agents real-life usage scenario

The Golden Agents project looks at the interaction between producers and consumers in various creative industries in Amsterdam during the seventeenth and eighteenth century, of which book production is an important one. As a case study we focus on a dataset of Occasional Poetry [=Gelegenheidsgedichten] (D_{OP})[3] that contains metadata on poems that are among others written to celebrate a marriage. Every poem in this dataset comes with the name of the bride and groom, the marriage date, and common bibliographic data such as information on the author and publisher. To enrich D_{OP} for a broader view and understanding of the people being chanted on (i.e. to get insight in their social network, or in their lifespan in general), D_{OP} is linked to

[1] https://lenticularlens.org/docs/03.Ontology/

[2] https://www.goldenagents.org/

[3] https://www.kb.nl/bronnen-zoekwijzers/kb-collecties/oude-drukken-tot-1801/gelegenheidsgedichten-16de-18de-eeuw

two (external) datasets of the Amsterdam City Archives: (i) The notice of marriage registries [=Ondertrouwregisters] (D_M)[4] that mentions a bride, groom and date of a notice of marriage and (ii) to the Baptism registries [=Doopregisters] (D_B)[5], which report on the baptism of children born out of a marriage by documenting the child's mother, father, witnesses and the date of the baptism.

3.2 Data Partition

The partitioning of a dataset is the process that enables users to define a data-subset on which one wants to match based on explicit restrictions on the type and optionally attribute-values of the entities to undergo entity matching.

Example 1 Figure 2 shows a partition of D_M. It restricts its selection to instances of `schema:Person` (entity-type) that are the subject of a poem describing an event containing the `rdfs:label` "marriage" (property-value) as the dataset distinguishes persons in the role of author from persons in the role of being chanted on.

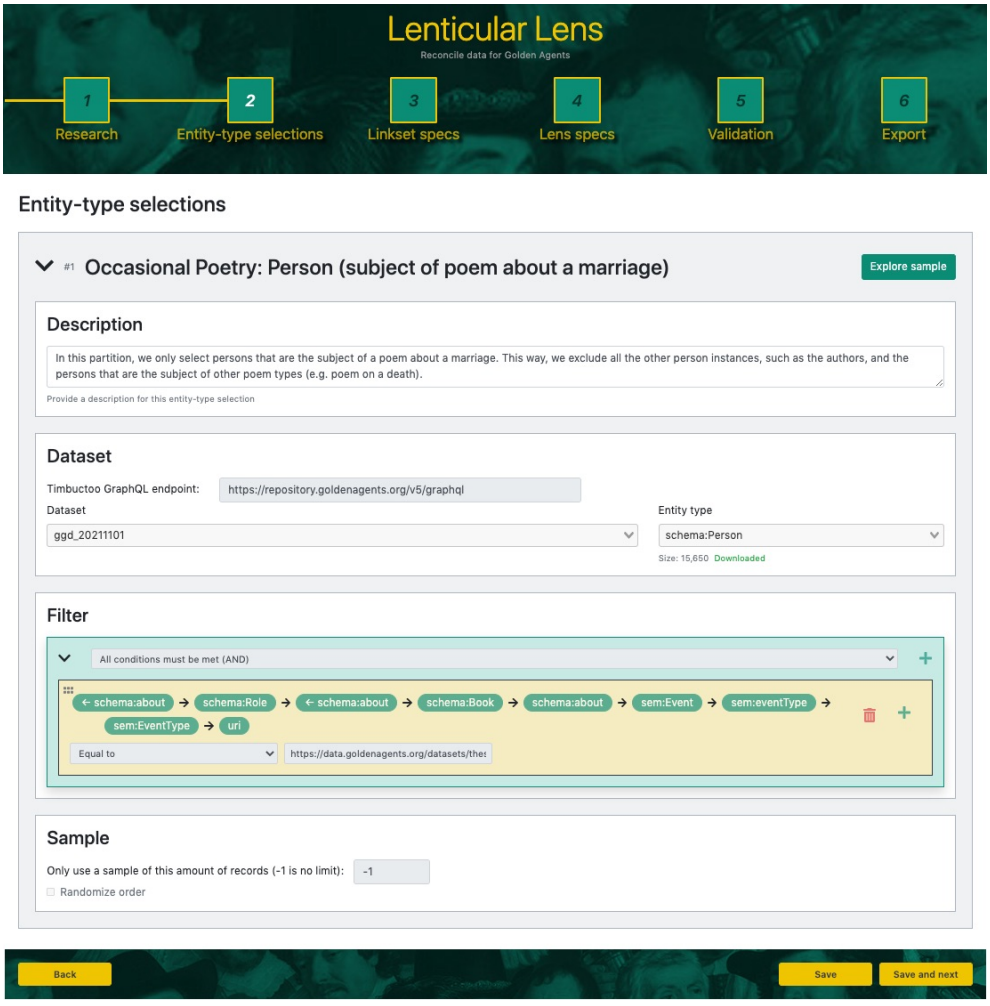

Figure 2: Partitioning of a dataset for entity matching.

4 https://archief.amsterdam/uitleg/indexen/45-ondertrouwregisters-1565-1811
5 https://archief.amsterdam/uitleg/indexen/57-doopregisters-voor-1811

3.3 Linkset: data linking

Once explicit partition-declarations on which one wants to match entities are specified for all datasets of interest, one can now specify how entities stemmed from a partition within a source-collection are to be linked to any entity of a partition stemmed from a target-collection. For this, users are required to indicate the attributes (identity criteria) of the selected entities over which a matching algorithm of their choice should be executed. In the event that more than one algorithm is needed for various type of comparisons, the user is to specify how they are to be combined using standard or fuzzy logic operators. The combination definition is then used to compute a final identity score for each discovered link.

The tool offers a variety of in-house, ad-hoc and off-the-shelf algorithms. The latter includes algorithms such as Levenshtein[6], Jaro and Jaro-Winkler[7], Soundex[8], Metaphone[9] and Double Metaphone.

Example 2 The linkset specification depicted in Figure 3 for link discovery between D_{OP} and D_B illustrates the invocation of the Levenshtein algorithm twice (name of the person and its partner) and the time delta computation once (for baptism events within 10 years of the marriage). It also shows that, as these matching methods are combined using a logic operator AND, for the discovery of a link in this setting, all methods' conditions are to be met.

3.4 Lens: link manipulation

It is fairly reasonable that in some cases more than a single set of links is useful for investigating a problem at hand. Sometimes, these sets intersect or complement each-other, or are simply relative complements. For dealing with the manipulation of sets of links, the tool offers set-like operators such as union, intersection, difference, and symmetric difference for the creation of a lens. Contrarily to the other operators, the intersection is not only applicable to links but also to sets of linked-resources.

Here too, the manipulation of sets can benefit from the use of means such as fuzzy logic for the final computation of a link's score depending on the operator used. For example, union and intersection can be implemented with standard or fuzzy logic. However, in the case of a (symmetric) difference the resulting links keep their original scores.

Example 3 In the Occasional Poetry use case, 4 sets of links are created such that each one of them can be studied separately. However, the scenario depicted in Figure 4 illustrates the particular need for grouping (1) couples with the intention of marriage (D_M) who got married (D_{OP}) within 6 months with (2) those celebrating at most 50 years of marriage anniversary (D_{OP}).

3.5 Link Evaluation

Data-driven investigations do not always have the same quality of input-data and do not all require the same level of result quality. For some, erroneous links do not have

[6] https://en.wikipedia.org/wiki/Levenshtein_distance
[7] https://en.wikipedia.org/wiki/Jaro-Winkler_distance
[8] https://en.wikipedia.org/wiki/Soundex
[9] https://en.wikipedia.org/wiki/Metaphone

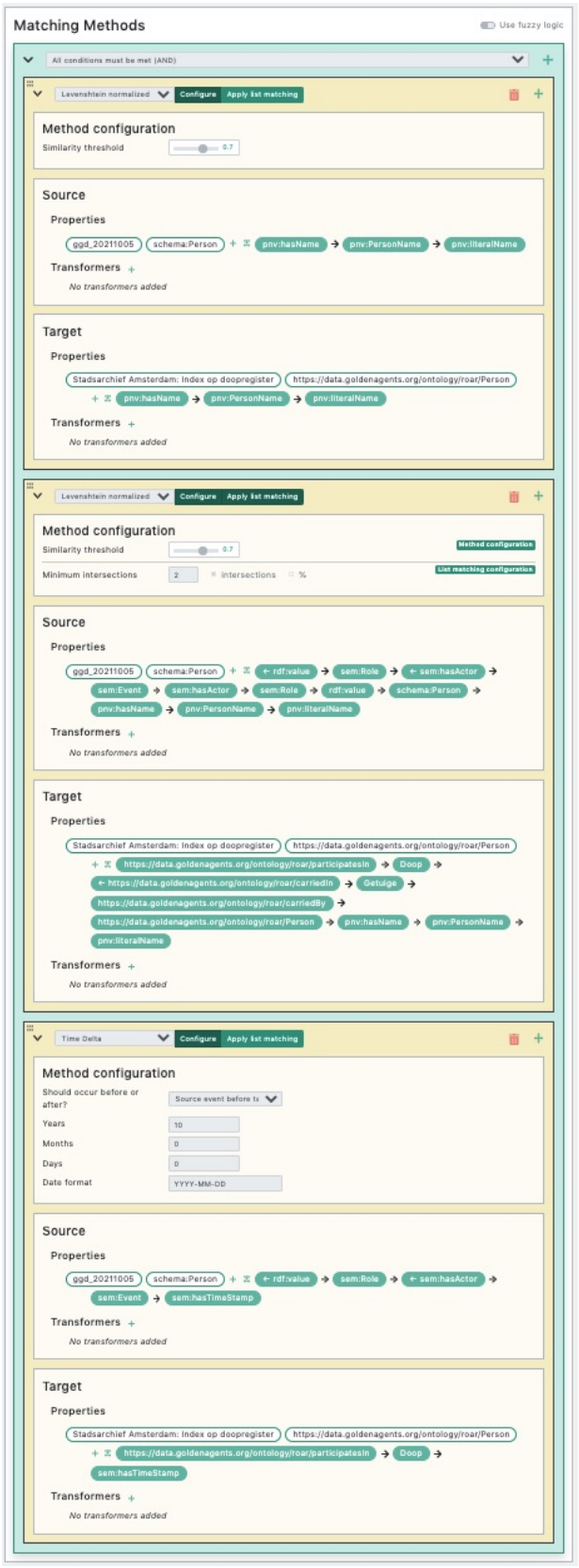

Figure 3: A specification for link discovery.

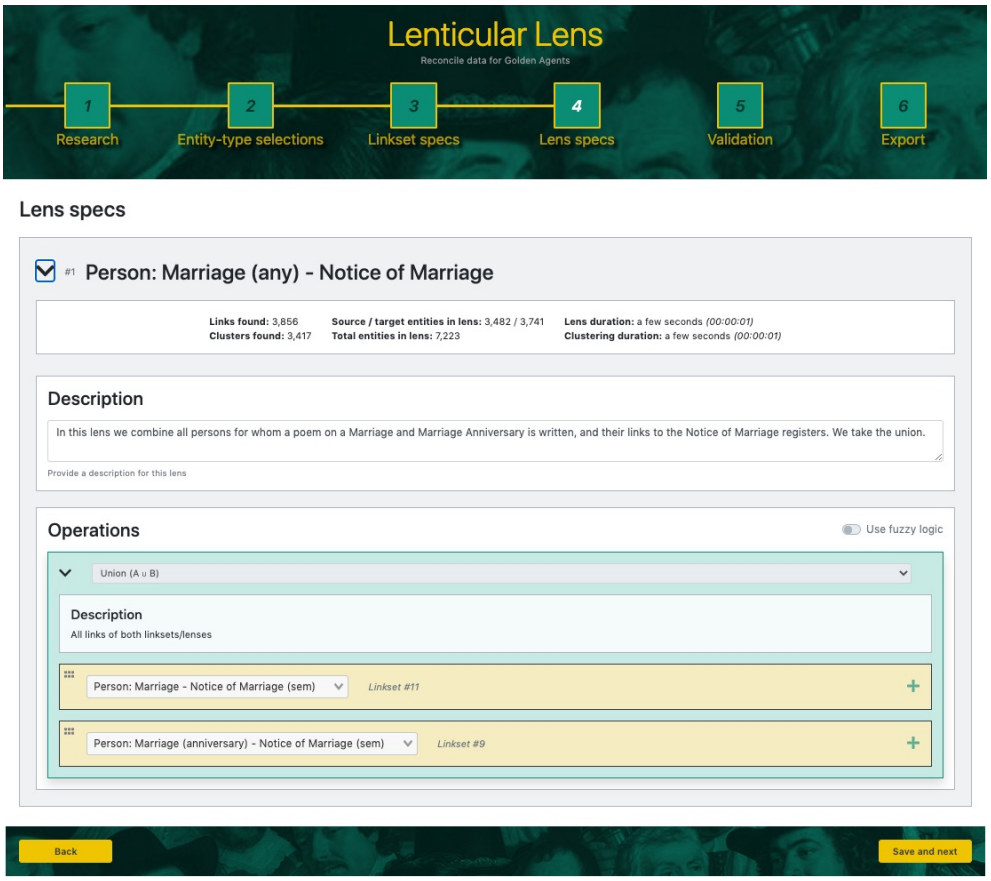

Figure 4: A lens specified as the difference between linksets.

a dramatic impact. However, for others, top quality links are of great importance. In the latter scenario for example, link evaluation is an option for only selecting perfect links, assessing the quality of the matching approach or the identity criteria. Here too, the tool allows users to assess links of their choosing and to document their assessments.

3.5.1 Evaluation Aid

Yes, the tool offers a basic "validation per link" feature. However, other means are at the users' disposal to still help them in time efficiency and consistency. They include various group-validation options and means for a compact visualization.

Cluster-based validations Once a set is created (linkset or lens), it is possible to cluster its links to provide users with co-referents. This allows them to then investigate (supposedly) related links together rather than apart.

Strength-based validations Users can also validate a group of links based on their respective strength. For example, one can decide to accept all links with a score above 0.94.

Visualizing Identity Link Network A network of co-referents groups various digital representations of the same real object. Depending on the quality of variants like

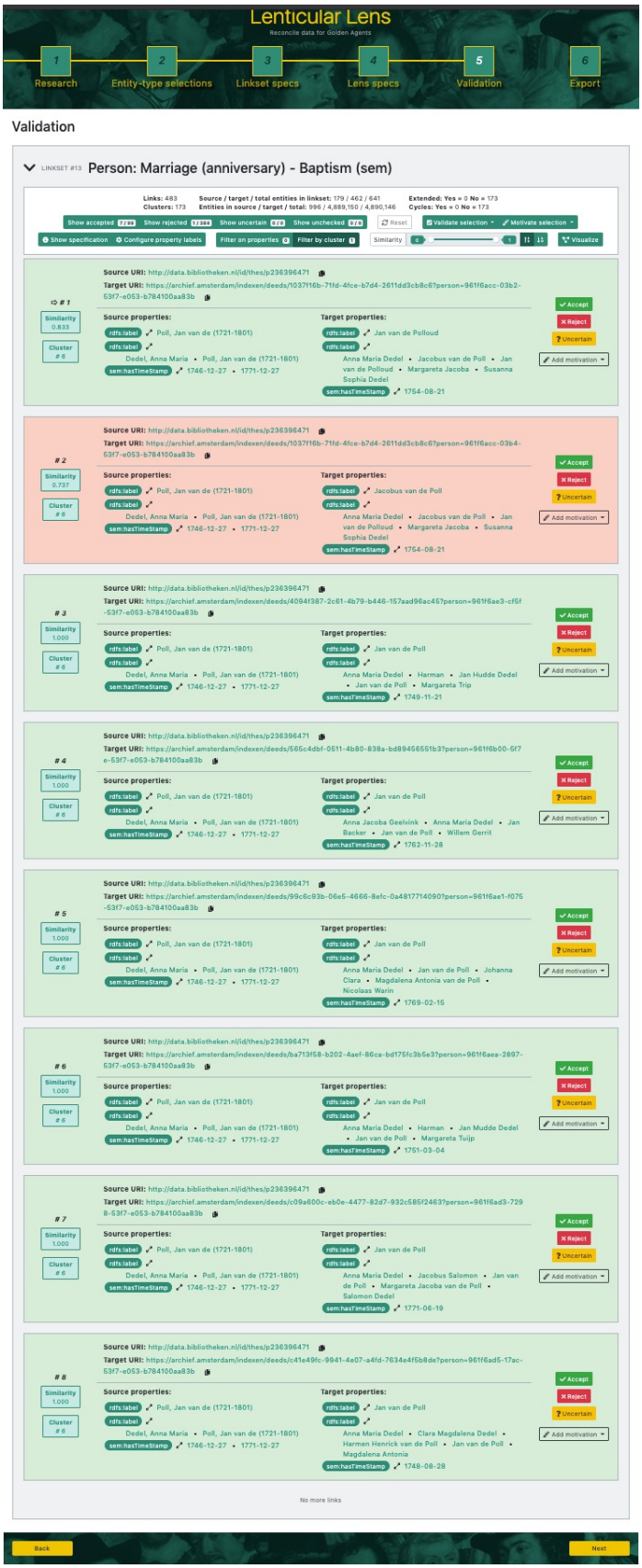

Figure 5: Overview of a validated cluster.

input-data, identity criteria and matching algorithms, the network varies in size. It grows bigger as a variant worsens and, the bigger the more tedious it is to render (readability and speed). Nonetheless, to help steer the validation process base on a meaningful network structure, the tool displays aggregated networks based on the similarity of the links' score.

Example 4 Figure 5 illustrates a validated cluster of nine co-referents where the red color indicates the only matching disagreement with the human validator out of the eight established links. An aggregation of the network based on the similarity of their connection (Figure 6) shows that seven resources share a link connection scored 1 while the remaining two are connected to the rest with lower scores (0.73 and 0.83). One goal is to support spotting weaker links, which have a reasonable chance of containing matching disagreements, provided strong discriminating criteria are available. It also facilitates the visualization of much bigger networks, such as the example from Figure 7, extracted from sameas.cc (Beek et al., 2018), composed of 439 nodes and 7,614 links.

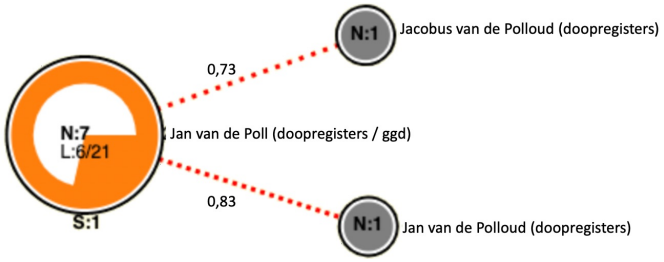

Figure 6: A compact identity network of Jan van de Poll.

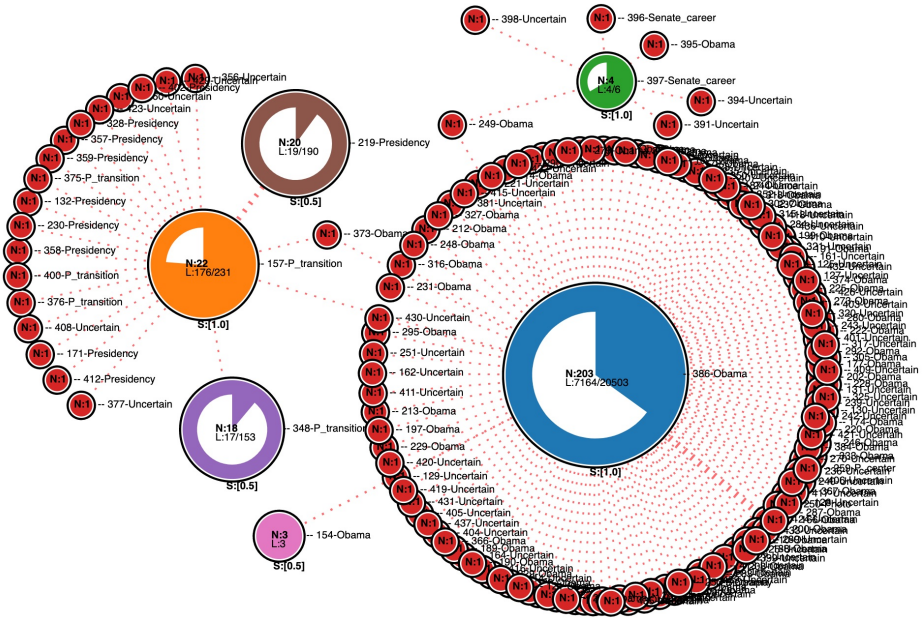

Figure 7: A compact identity network of Obama. It is originally a network of 439 nodes and 7,614 links.

3.6 Data Export

As shown in Figure 8, the tool offers an option to export linksets and lenses with their respective metadata of choice. By default, one can expect the matching data to be exported in RDF format with all its metadata attached using the standard reification approach while linking matched resources with the `owl:sameAs` predicate. Alternatively, the user can opt to not use reification at all or opt for other formats such as RDF* or singleton. Also, an alternative link predicate substitute such as well known `skos:broadMatch, skos:closeMatch, skos:exactMatch, skos:narrowMatch` and `skos:relatedMatch` or a custom predicate is possible, as well as exporting a CSV format.

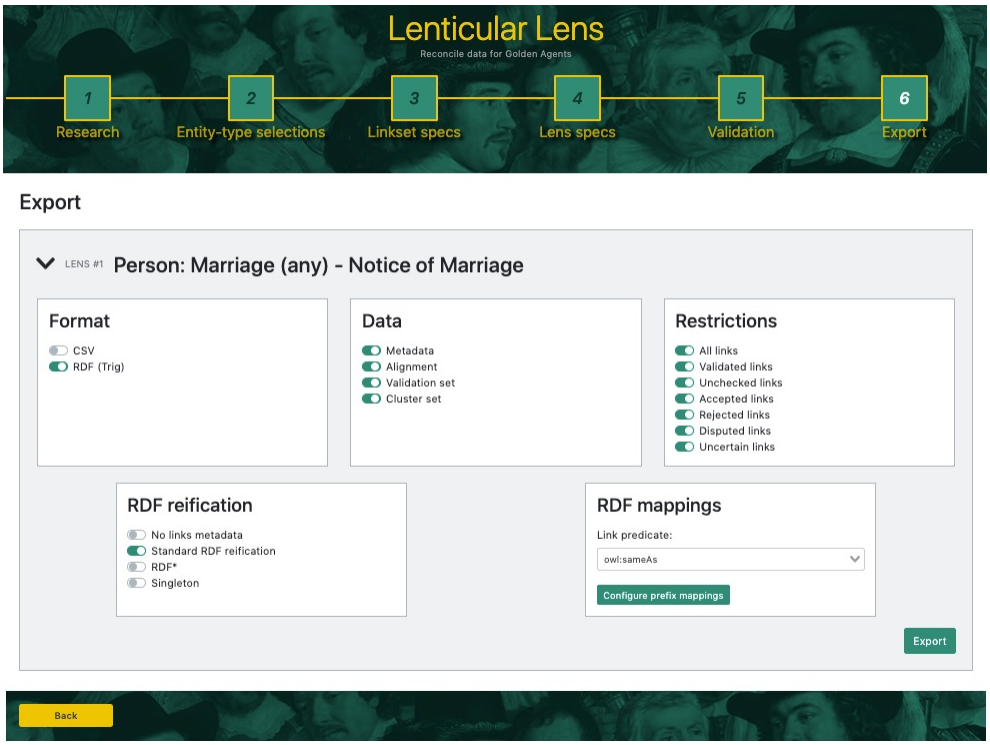

Figure 8: Options for exporting matching data and metadata.

4 Related Work

Lenticular Lens can be compared to the matching functionality of OpenRefine (Delpeuch et al., 2021). Although OpenRefine appears as an appealing tool for Humanities scholars for data harmonization or black-box entity matching, it lacks the flexibility, transparency and broad capability offered by the Lenticular Lens.

SILK (Volz et al., 2009b) is another tool similar to the Lenticular Lens. It utilises SPARQL protocol, caching and indexing techniques to access data. Like Lenticular Lens, it enables the combination of off-the-shelf algorithms and the combination of matching scores. However, the Lenticular Lens offers more flexible and expressive ways for such combination, from nested logic 'boxes' to fuzzy logic operators. Moreover, while SILK employs a dedicated XML-based language specification for metadata export, the Lenticular Lens offers RDF-based metadata inspired from VoID, a shared vocabulary, besides aid for the manipulation and validation of links.

5 Conclusion

We present a multi-purpose environment[10] to address in a flexible and generic way the ubiquitous issue of entity disambiguation through various facets of links: Creation – where the use of generic ad-hoc and off-the-shelf methods can be logically combined for the discovery of matching resources and computation of their identity confidence-scores; Manipulation – where set-like operators allow for combining sets of links (linksets and/or lenses); Validation – allows for assessing discovered links and/or matching methods using a basic approach (one link per validation) or aids such as bulk validation and aggregated visualization based on identity confidence-score; Documentation – allows for explicitly documenting all specifications required by the tool, thereby facilitating reproducibility, assessment, and re-use.

The Lenticular Lens has been successfully used to address several Humanities use-cases within the Golden Agents project thanks to its genericity and flexibility. We have integrated 17^{th} and 18^{th} centuries resources, some of which, regarding occasional poetry, are presented in this paper. They enable the project to gained insight in the social circles of notable persons and the communities they participated in. This opens doors for a broader look into the production of culture and helps to understand to what extent cultural industries interact with each other professionally and socially (e.g. are people connected through confession, or family bonds).

For future work, we plan to incorporate the equality metrics for an identity network (Idrissou et al., 2020) as well as Reconciliation (Idrissou et al., 2019) as an aid to validation, besides more off-the-shelf matching and clustering techniques. Another interesting path would be to investigate how to join efforts with similar approaches such as SILK.

Acknowledgement

We thank Kerim Meijer for his hard work on the development of the tool and fruitful comments, and Chiara Latronico for her dedication for helping to achieve a user friendly interface.

References

Alexander, K., R. Cyganiak, M. Hausenblas, and J. Zhao (2009). Describing linked datasets. In LDOW.

Alexander, K., R. Cyganiak, M. Hausenblas, and J. Zhao (2011). Describing linked datasets with the void vocabulary.

Beek, W., J. Raad, J. Wielemaker, and F. Van Harmelen (2018). sameas. cc: The closure of 500m owl: sameas statements. In European semantic web conference, pp. 65–80. Springer.

Delpeuch, A., T. Morris, D. Huynh, S. Mazzocchi, Jacky, W. (bot), T. Guidry, O. Stephens, I. Matsunami, I. Sproat, S. Santos, allanaaa, K. Trivedi, E. Mishra, M. Magdinier, L. Liu, J. Ong, F. Tacchelli, F. Giroud, A. Nordhøy, A. Beaubien,

[10] Users are invited to try the tool. The source-code, live-tool and documentation can be found at https://github.com/knaw-huc/lenticular-lens, https://lenticularlens.goldenagents.org/ and http://lenticularlens.org/ respectively.

M. Saby, L. Chandra, B. Lehečka, R. Fontenelle, Y. Shahrabani, noamoss, and xseris (2021, November). Openrefine/openrefine: Openrefine v3.5.0.

Hassanzadeh, O., A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang (2009). A framework for semantic link discovery over relational data. In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 1027–1036.

Hassanzadeh, O., R. Xin, R. J. Miller, A. Kementsietsidis, L. Lim, and M. Wang (2009). Linkage query writer. In Proceedings of the VLDB Endowment, Volume 2, pp. 1590–1593.

Idrissou, A., F. van Harmelen, and P. van den Besselaar (2020, nov). Network metrics for assessing the quality of entity resolution between multiple datasets1. Semantic Web 12(1), 21–40.

Idrissou, A., V. Zamborlini, C. Latronico, F. van Harmelen, and C. van den Heuvel (2018). Amsterdamers from the golden age to the information age via lenticular lenses. In DH Benelux Conference 6-8 June 2018, International Institute for Social History, Amsterdam.

Idrissou, A., V. Zamborlini, F. Van Harmelen, and C. Latronico (2019, sep). Contextual Entity Disambiguation in Domains with Weak Identity Criteria. In Proceedings of the 10th International Conference on Knowledge Capture, New York, NY, USA, pp. 259–262. ACM.

Mirani, A. H. and M. A. Radke (2014). Graph based disambiguation of named entities using linked data. International Journal on Recent and Innovation Trends in Computing and Communication 5(6), 78–86.

Ngomo, A.-C. N. and S. Auer (2011). Limes—a time-efficient approach for large-scale link discovery on the web of data. In Twenty-Second International Joint Conference on Artificial Intelligence.

Volz, J., C. Bizer, M. Gaedke, and G. Kobilarov (2009a). Discovering and maintaining links on the web of data. In International Semantic Web Conference, pp. 650–665. Springer.

Volz, J., C. Bizer, M. Gaedke, and G. Kobilarov (2009b). Silk-a link discovery framework for the web of data. In Ldow.