Documenting the Creation, Manipulation and Evaluation of Links for Reuse and Reproducibility

Idrissou, Al; Zamborlini, Veruska; Kuhn, Tobias

published in
Knowledge Engineering and Knowledge Management
2022

DOI (link to publisher)
10.1007/978-3-031-17105-5_6

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

Documenting the Creation, Manipulation and Evaluation of Links for Reuse and Reproducibility

Al Idrissou[1](✉), Veruska Zamborlini[2], and Tobias Kuhn[1]

[1] Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
{oid201,t.kuhn}@vu.nl
[2] Federal University of Espirito Santo, Vitoria, ES, Brazil
veruska.zamborlini@ufes.br

Abstract. Data integration is an essential task in the open world of the Semantic Web. Many approaches have been proposed that achieve such integration by linking related entities across data providers, but they lack the support for in-depth documentation of the involved processes such as the creation, manipulation and evaluation of links. As a consequence, detailed documentation that eases the understanding and reproducibility of underlying processes is needed for a reliable reuse of graphs of identity available in the Semantic Web. We present here an approach to document such links and their processes, building upon a representation we call VoID+. It enables link-publishers to provide data-users with information that better support them in accessing and using links. We show that our approach with the proposed VoID+ ontology allows us to address the relevant competency questions around the reuse of integrated Semantic Web data. We also demonstrate how our approach has been successfully implemented in the Lenticular Lens, a user interface tool that annotates links it discovers, manipulates or validates under user's guidance. Based on a real-life humanities case study, we can show that the ontology amply annotates links in its life-cycle for reliable decision making by data-users.

Keywords: semantic web ontology · semantic web vocabulary · ontology design · data integration · linkset

1 Introduction

Links connecting co-referent entities (a.k.a identity links[1]) constitute one of the pillars of Data Integration in the Semantic Web where entity matching techniques enable their discovery and *creation*. They are represented as RDF triples where the subject and object are described in one or more datasets. They can be grouped together in sets (or "graphs") and can be annotated with metadata such as matching scores to enable seamless navigation across datasets and hence increase the potential of addressing complex problems such as investigating innovation on the creative industries during the long Dutch Golden Age (Sect. 4.2).

[1] For readability, we use the terms "links" and "identity links" interchangeably.

O. Corcho et al. (Eds.): EKAW 2022, LNAI 13514, pp. 81–96, 2022.
https://doi.org/10.1007/978-3-031-17105-5_6

As the quality of input-data, matching algorithms and discriminating criteria are not always perfect and the matching process is heuristic in nature, erroneous links may be introduced, forcing data-users to *evaluate* the links' quality prior to their usage. More importantly, for *reliable reuse*, one is expected to *assess* whether the context at hand fits the context in which selected links are discovered. A given link may originate from (i) the application of given methods under particular data filters and matching criteria or from (ii) *manipulation* (combination, intersection, ...) of other existing links created by different providers. At present, the context of the origins of links lacks a uniform and coherent method of representation.

Highlighting what was mentioned above, integrating data via links depends not only on the creation and/or manipulation of links but more so on their quality and the ability to understand their provenance. This motivates our argument that the potential for a set of links to qualitatively complement an interlinked dataset depends on the quality of its links which, in its turn, depends on the quality of the processes leading to their creation. Evaluating these processes starts with accessing the *metadata*, which is often lacking or not comprehensible enough. Indeed, most links in the Semantic Web are provided without much provenance, often embedded within datasets or as 'plain' sets of 'owl:sameAs' or 'skos:closeMatch'-like links [3]. Moreover, works meant for guiding or assessing link creation do not highlight the need for documenting it [1]. Ignoring this is an indulgence that may not always be worth it. Instead, we advocate that the reliance on metadata is of paramount importance to enable graphs of links to be reliably assessed, evaluated, reproduced, reused or queried in the search for interlinked datasets. So much so that [2] writes: "it is in the best interest of a dataset publisher to provide potential users of the data with information that supports them in accessing and using the dataset". Based on this, we investigate the following questions: *"how can we detail the documentation of the creation, manipulation and evaluation of links and hence enable the understanding and reproducibility of interlinked datasets so that they can be reliably (re)used?"*.

This paper aims to provide a means to a comprehensive semantic documentation of links and their processes. As such, the main contributions of this work are: (1) A conceptual overview of key elements to comprehensibly describe links and their processes; (2) VoID+, a concrete representation for the proposed elements that goes beyond the documentation of interlinked datasets; (3) in-depth insights into the effects of the proposed approach on a real-world use case.

Sections 2 and 3 present the state-of-the-art and the proposed approach. The latter is evaluated in Sect. 4 by proposing SPARQL queries addressing the competency questions, and presenting the results of their application to a real case study. The use case was developed and executed by humanities researchers in the Golden Agents project using the Lenticular Lens[2] tool, briefly introduced in Sect. 4, which implements VoID+, the proposed vocabulary. Section 5 presents some points of discussion while Sect. 6 concludes the article.

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

1.1 Competency Questions

At first glance, when a user comes across a set of discovered links, there are typically a number of nagging questions one wishes to answer for a reliable re-use, evaluation or reproduction. For example, one would like to know which sources and entities are covered, what algorithms and discriminating criteria are used, if the links are validated or if the resources are clustered. We propose here some competency questions for which a set of links should provide answers:

1. Given a set of links (Linkset) of interest. **(a)** What are the interlinked datasets involved? **(b)** If any, what sequences of restrictions (entity-types), property selections and value filter are applied to the entities of interest and how are the elements of a sequence of restriction combined? **(c)** What entity matching techniques (algorithms) are applied? If more than one is applied, how are they combined? **(d)** For a particular matching method, on which resource descriptions (property-values) are they applied; under which value-constraints and threshold?
2. Given a set of datasets and entity-types, what links are returned to the user if she is only interested in the ones that are: **(a)** Above a certain threshold? **(b)** Found by a specific method? **(c)** Validated as accepted? **(d)** Rejected above a certain threshold?
3. What set(s) of links is/are returned to a user interested in: **(a)** A certain dataset and/or entity type? **(b)** The use of a particular algorithm for link discovery? **(c)** A set of discriminating properties?
4. Given a set of links composed of multiple Linksets of interest: **(a)** What operators are used to generate the set of Linksets? **(b)** How are the operands combined? **(c)** How are each of the operands generated?

Few of the above questions can be answered using existing vocabularies but with limitations, for Example, 1 a/b and 2.b can be addressed using VoID [2] and/or Prov-O[3]. However, the level of details required to properly address each question is not achievable with current approaches.

1.2 Motivation

To support a comprehensible documentation of discovered links and their processes so that questions such as the ones above can be addressed, we highlight some motivating concepts area we intend to challenge. These are:

- **Data partitioning:** entities selected for matching purposes often do not constitute the whole data at hand, but a subset based on a specific type of entities but also on particular properties and their values. For example, only selecting from a university dataset AI students who had an internship.
- **Identity criteria:** the identity-link discovery process requires a set of rules or criteria for enabling the discovering links. These should be made explicit.

[3] https://www.w3.org/TR/prov-o/.

- **Multiple data sources:** a set of links may point to multiple data sources. Conceptually, nothing stops the source or target to be composed of multiple sources, contrary to what is observed in some vocabularies (see Sect. 2).
- **Multiple linking algorithms:** often, more than one type of metrics are used for object description comparison [6] only, no representation explicitly elaborates on their logical combination for reaching a combined link score.
- **Link-dataset manipulation:** as links can be produced under different settings/providers, the ability to manipulate those sets of links and their respectively computed matching scores using set-like operators may be informed.
- **Clustering:** the purpose of integration is to group co-referents[4], instances of heterogeneous sources, to address one or various pressing life problems.
- **Validation:** the reporting on the evaluation of not only the quality of links but also the quality of link processes enabling the mapping of digital representations to their unique counterpart in the sphere of real world objects.

To the best of our knowledge, at present, in the Semantic Web, no existing approach or vocabulary provides such detailed and broad coverage provenance on the integral processes surrounding the discovery of the links, their combination, and on whether and how they have been manipulated, clustered or evaluated.

2 Related Work

This section presents the related work on the VoID vocabulary and its limitations, on a number of approaches for representing links and their related processes, and on vocabularies used in the Semantic Web data integration tools.

2.1 VoID

The Vocabulary of Interlinked Datasets (VoID) [2], a standard advised by W3C, is a general purpose core vocabulary for providing metadata on datasets including graphs of links for the discovery and usage of interlinked datasets. However, it is not designed to dive into granular descriptions of specific datasets such as those only composed of links. Consequently, it is unable to inform data-users on *how links are generated and how the target datasets are semantically related.*

The observation is that it is (i) *a source of potential ambiguity in describing a partition* as it falls short of differentiating between a partition formed of entities that are for example [AIStudents and (had an internship or had an exchange program)] from one formed by entities that are [AIStudents or had an internship or had an exchange program] or in providing means to *restrict entities based on the value-range* of a property, for example, students born before the year 2000. Furthermore, for real world problems, it has a (ii) *too strict definition for a graph of links.* It forces instances of `void:Linkset` to be *directed* and to hold between *exactly two non-identical datasets* thereby explicitly disallowing a set of links to

[4] Co-referent is a term used in entity matching jargon to indicate a set of resources pointing to the same real-life object.

have more than one `void:subjectsTarget` or `void:objectsTarget`. This semantic does not allow, for example, for a linkset to hold for a dataset deduplication as it implies establishing links between duplicated resources stemmed from the *same dataset* with the intention of removing/merging redundant data.

2.2 VoID Extensions

The VoID ontology has seen a number of extensions over the years. **VoIDext** [5] is a vocabulary designed to enable the documentation of federated SPARQL queries in a way that highlights relatedness between datasets such that machines and humans can benefit. It proposes the concept *Virtual Links*, extending VoID with respect to querying links rather than detailing instance matching. **VoIDgen** [4] is designed for automating the description of large datasets using VoID by applying MapReduce paradigm to discover (sub)datasets. Besides reducing manual effort, incompleteness and inaccuracy, it proposes concepts such as Crisp versus Fuzzy linkset, enriching the semantics of datasets. **VoIDp** [12], is an ontology designed for the enhancement of interoperable datasets with virtual links.

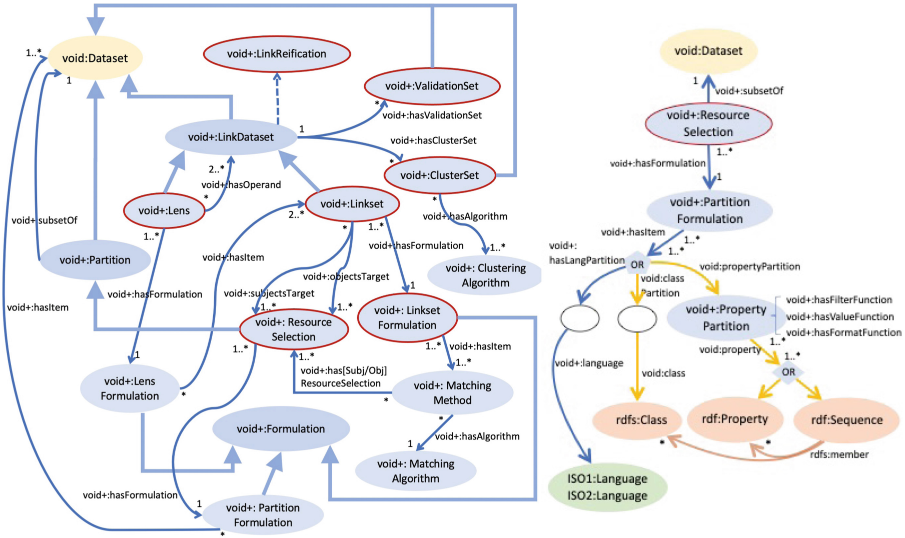
Overall, the existence of extensions such as VoIDp, VoIDext and VoIDgen illustrate three interesting points crucial for the maturity of interlinked datasets. First, the extensions show the conformity with best practice, which advocates the *reuse of well-known vocabularies wherever possible*. Second, they exhibit the *acceptance of VoID* as a standardised core vocabulary for annotating interconnected datasets. Last but not least, they yet reveal limitations of VoID, hence the *need for new concepts* that best tackle the respective domains being modelled.

2.3 Vocabulary Used in Data Integration Tools

To a certain extent the metadata reporting on the links' provenance is addressed by some matching approaches / frameworks. SILK [15] provides a number XML-based files containing the matching specification. Only the resulting links are provided using RDF format, though it does not follow a known reification format for commenting on the identity links. It provides some means to use more than one matching method and combine the resulting scores using operators such as MAXIMUM, MINIMUM and AVERAGE. LIMES [10], another matching framework, have recently provided an RDF vocabulary called LIMES Configuration Ontology (LCO). However, as the name suggests, its main purpose is to express LIMES' linking-configuration in RDF. In the process, it uses VoID, but only specializing its main class, void:Dataset. It also provides a means to use more than one matching method and combines the resulting scores using operators such as AND, OR, MINUS and XOR. To the best of our knowledge, other approaches do not offer better means of documentation.

3 Approach: VoID+

The ontology here presented is called VoID+ as it is meant to extend and be compatible with the VoID vocabulary. Figure 1a provides a simplified overview of



(a) The VoID+ ontology and its main concepts.

(b) Selecting a matching resource in VoID+.

Fig. 1. Linkset's method and context in VoID+

VoID+[5] highlighting (in bold-red-outline) its main elements: *Resource Selection*, *Matching Method*, *Lens*, *Validation-set* and *Cluster-set*. In the sequel, these main elements and their properties are further explained.

Resource Selection. This aspect concerns the selection of the resources under scrutiny that can potentially end up being determined to be co-referent entities during an entity matching process. Therefore, to perform a matching, one first needs not only to select one or more data-sources, but also to restrict which resources within each source will undergo the matching. The first way of doing so is by applying a type restriction. Down this line, further restrictions can be applied by forcing the value of a number of properties to lie within a certain range. A Resource Selection is thereby the annotation of such a selection process.

In the excerpt depicted in Fig. 1b, for entity selection purposes we propose the entity type `voidPlus:ResourceSelection`, which is a `voidPlus:Partition` based on a `void:classPartition` and/or a `void:propertyPartition`. While the relation `void:classPartition` solely consists in specifying the type of entity under scrutiny, the `void:propertyPartition` entails a little more. It consists in specifying a property or property path and a restriction that the selected property should undergo for the selection of the right entities for the further down the road entity matching process. Those restrictions can be combined using a formula description given by `voidPlus:hasFormulaDescription`.

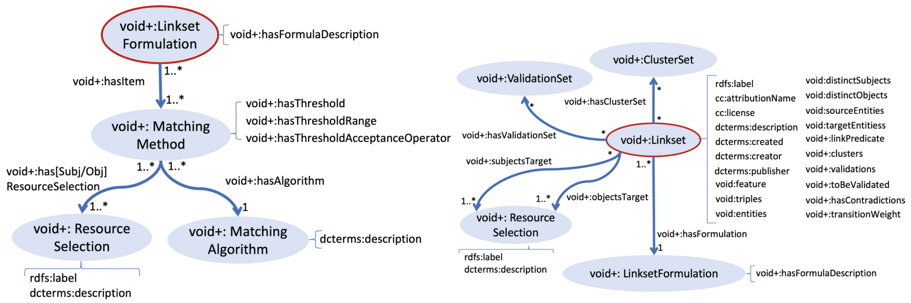
[5] Details at https://lenticularlens.org/docs/03.Ontology/ and https://tinyurl.com/VoIDPlusGit.

Linkset Formulation For simple matching problems, finding co-referents can be done using a single matching algorithm. However, more than one is often needed for practical reasons. In this latter scenario, clearly reporting on how they work together for detecting co-referents is essential. As depicted in Fig. 2a, a Linkset Formulation entity is a resource for just doing the aforementioned.

Once resources of type Resource Selection are created, one can go ahead and use them for specifying the restricted collections to be used in a particular Matching Method. A resource of type `voidPlus:MatchingMethod` then specifies the Matching Algorithm and its arguments such as threshold, range and operator. In the end, all Matching Methods used in a matching process are documented using a resource of type `voidPlus:LinksetFormulation` which explicitly documents how they bind together in a logic expression given by the predicates `voidPlus:hasFormulaDescription` and `voidPlus:hasFormulaTree`.

Linkset. Linkset metadata (Fig. 2b) includes the WHO - WHAT - WHEN - HOW and related processes explaining the aboutness of links. While a Resource Selection entity specifies WHAT to match as subject and object targets, a Linkset Formulation specifies HOW entities are matched. Furthermore, some statistics on the matching results and other information can be reported such as the number of links found, the numbers of entities linked, WHO created the linkset and WHEN.

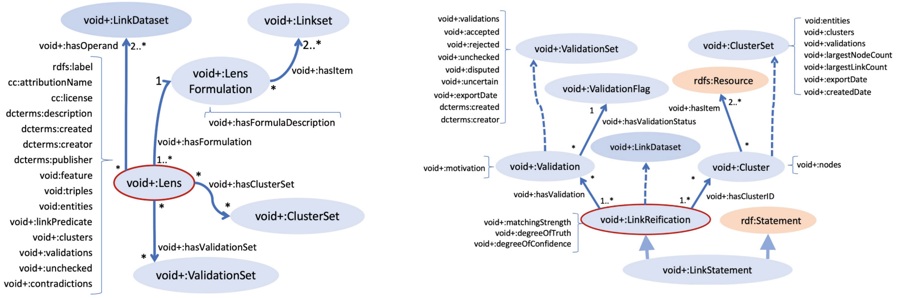
As discussed earlier in this section, according to the VoID documentation, the `void:Linkset` definition expects as data-sources exactly one source and one target, different from each other. This means it is more restrictive than the `voidPlus:Linkset` proposed here, since the latter also allows a linkset to connect resources *within* a single data-source or across more than two. As a consequence, we define a new concept rather than reusing `void:Linkset` in an incompatible manner. (`void:Linkset` is also not a subclass of `voidPlus:Linkset` as the latter requires the description of the processes underlying the creation of the links.)



(a) Specifying the way in which methods are logically combined in VoID+.

(b) Specifying a linkset's context in VoID+.

Fig. 2. Linkset's method and context in VoID+



(a) Specifying the Lens context in VoID+.

(b) Specifying the Reification of Links in VoID+.

Fig. 3. Specification of a lens and a link Reification in VoID+

Lens. The creation of Lenses is another process that is important to document. In short, a lens is the result of a set-like operation — UNION, INTERSECTION or (SYMMETRIC) DIFFERENCE — over one or more Linksets and/or Lenses. For that, the entity `voidPlus:Lens` documents (Fig. 3a) its constituents as `voidPlus:hasOperands`. Like a `voidPlus:Linkset`, a `voidPlus:Lens` metadata includes `voidPlus:LensFormulation`, among others, and points to resources of types `voidPlus:Clusterset` and a `voidPlus:Validationset` if available.

LinkReification. The `voidPlus:LinkReification` resource, depicted in Fig. 3b, represents a reified link as a possible content of a `voidPlus:LinkDataset`. When reified, for example as an rdf:Statement using a standard reification or any other reification approach [7,9,11,13] for the matter, one can then annotate it with its own properties. For each link, VoID+ allows for one or more validations (`voidPlus:hasValidation`) by one or more users, belonging to one or more clusters (`voidPlus:hasClusterID`) created by different algorithms/processes. VoID+ does not enforce a particular reification type and link predicate, giving freedom of choice to the links creators. It requires, however, the users to be aware of the creator's choices when reusing or querying the linksets.

ClusterSet. One or more `voidPlus:ClusterSets`, using clustering algorithms, can be provided for a linkset or lens, which form a `voidPlus:LinkDataset`. This can support the validation process allowing for an overview of potentially equivalent entities.

ValidationSet: Qualitative Evaluation. When available, instantiations of one or more `void+:ValidationSet` is attached to a linkset or lens (`void+:LinkDataset`), comprising metadata with statistics and authority information on the validation process. Statistics on this matter can be included in the linkset metadata, particularly including eventual contradictions when more than one validation is provided, where for example some validations present a link as correct while other validation statements flag it differently.

4 Evaluation

The proposed vocabulary is evaluated by its ability to answer the proposed competency questions (Sect. 4.1), which are implemented as SPARQL queries and applied to a real case study. Due to space limitation, we only present two queries and their results while the remaining five queries and results are available online[6]. At present, only Question 2 cannot be fully answered as Item 2.b particularly challenges the proposed representation, as discussed in Sect. 5. In general, the complexity of the queries ranges from simple queries (4 patterns) to complex ones (recovering complex property paths used is a matching).

4.1 Queries For Competency Questions (QCQ)

Matching Results described using VoID+ help unveiling and understanding the sequence of processes leading to the creation, manipulation, clustering and validation of discovered links. Hereby, we present the queries addressing questions 1a & 1b and 2a & 2d.

QCQ 1.a & 1.b Given a linkset of interest, the query presented in Listing 1.1 addresses part of question 1 by (a) retrieving interlinked datasets and (b) their explicit partitions. This exhibits the selected class, properties and/or languages and how they are combined in a formula.

```
SELECT DISTINCT ?PartitionLabel ?PartitionType
?Restriction ?PartitionInFormula ?filterFunction ?filterValue
{
  <--LINKSET--> voidPlus:subjectsTarget|voidPlus:objectsTarget ?rscSelection.

  # a) A dataset that is not itself a ResourceSelection
  ?rscSelection     voidPlus:subsetOf*               ?ds ;
                    rdfs:label                       ?PartitionLabel ;
                    voidPlus:hasFormulation          ?formulation .
  MINUS { ?ds a voidPlus:ResourceSelection . }

  ?formulation voidPlus:hasItem        ?partition  .
  OPTIONAL { ?formulation voidPlus:hasFormulaTree ?PartitionInFormula . }

  # b) Restrictions on the selected resources
  ?partition  a ?PartitionType;
  { ?partition  void:class | voidPlus:language | void:property  ?Restriction.
    FILTER (!isBlank(?Restriction)) }
  UNION { SELECT ?partition
    (GROUP_CONCAT(DISTINCT ?propidx; SEPARATOR=" \n ") AS ?Restriction)
    WHERE {
      ?partition void:property _:pseq .
      _:pseq   a      rdfs:Sequence ;
               ?seq   ?prop .
      FILTER (?seq != rdf:type )
      BIND(CONCAT(strafter(str(?seq),"_"), " " ,str(?prop)) as ?propidx)
    } GROUP BY ?partition  }
  OPTIONAL { ?partition  voidPlus:hasFilterFunction    ?filterFunction ;
                         voidPlus:hasValueFunction     ?filterValue .  }
} ORDER by ?PartitionLabel  ?Restriction
```

Listing 1.1. Interlinked datasets & filters (1.a and b)

[6] https://github.com/VoIDPlus-owl/EKAW2022/blob/main/CQs_for_VoIDPlus__EKAW.pdf.

QCQ 2.a & 2.d Given a set of data and entity-types of interest, Listing 1.2 retrieves links scored above 0.75 and rejected by a user in a validation process. Item (c) can be similarly addressed by selecting the accepted ones.

```
SELECT DISTINCT ?linkDataset ?subDs ?sub ?objDs ?obj ?strength
{
  VALUES ?givenDs { <--- SPECIFY THE DATASETS OF INTEREST ---> }
  VALUES ?givenType { <--- SPECIFY THE TYPES OF INTEREST ---> }

  ?linkDataset      voidPlus:hasOperand* /
    ( voidPlus:subjectsTarget | voidPlus:objectsTarget ) ?rscSelection .
  # Datasets Restrictions
  ?rscSelection     voidPlus:subsetOf+      ?givenDs ;
                    voidPlus:hasFormulation ?formulation .
  # Type Restrictions
  ?formulation      voidPlus:hasItem        ?typePartition  .
  ?typePartition    void:class              ?givenType .
  # Standard Linkset Reification
  graph ?linkDataset { ?link   rdf:subject                    ?sub ;
                               rdf:object                     ?obj ;
                               voidPlus:matchingStrength      ?strength . }
  # Finding the dataset/entity-type selections for ?subj and ?obj
  ?linkDataset  voidPlus:hasOperand* / voidPlus:subjectsTarget /
                voidPlus:subsetOf*   /   rdfs:label                ?subDs ;
                voidPlus:hasOperand* / voidPlus:objectsTarget /
                voidPlus:subsetOf*   /   rdfs:label                 ?objDs .
  # 3.a links above a certain threshold (0.75)
  FILTER (?strength > 0.75)
  # 3.d links that have been accepted
  graph ?linkDataset { ?link voidPlus:hasValidation        ?v }
  graph ?validationSet { ?v  voidPlus:hasValidationStatus   resource:Rejected}
}
```

Listing 1.2. Links rejected yet above a preset threshold of 0.75 given a set of datasources and entity-types of interest. (2.a and d)

4.2 Use Case - Occasional Poetry

VoID+ is in use via the Lenticular Lens, a data-integration tool developed in the context of several projects[7], namely RISIS, CLARIAH and Golden Agents. The use-case discussed here is run over the latest implementation of the tool within the Golden Agents project, which uses the tool for *integrating data to engage in the investigation of complex problems that span the interaction between productions and consumption of the creative industries during the long Dutch Golden Age.*

The Lenticular Lens. It is a flexible tool[8] that aims at utilising generic off-the-shelves algorithms and a few tailored ones to allow the discovery of links across multiple datasets through users' guidance. Using the proposed VoID+ ontology, it allows detailed documentation of user's tailored matching processes and the full or partial export of these documentations in various reification flavors at the user's convenience. Not only does it support the documentation of links discovery but it also supports those of link manipulations, clustering and validations.

[7] http://risis.eu/, https://www.clariah.nl/, https://www.goldenagents.org/.

[8] The tool and the present use case implementation can found at: https://lenticularlens.goldenagents.org/?job_id=90b598f72088ebd0e21446a12e353ffd.

Datasets. The Golden Agents project is working around the clock to make available to the public as many historical RDF datasets as possible by (i) converting humanities data into RDF and (ii) more than ever, enriching and linking the data of interest to the project. At present, it has in its Linked Open datasets portfolio a total of 27 and counting datasets of interest from 12 different content providers. The Lenticular Lens tool plays an important role since the project aims to encourage the addition of more and more RDF datasets to be interlinked, validated and reused by the public in the context or their own research.

As for the real life use case presented here, the following datasets are used:

- **SAA**[9]. The Amsterdam City Archives aka SAA, documents three social events that include *Trouw* (Marriage), *Doop* (Baptism) and *Begraaf* (Burial).
- **Occasional Poetry (Gelegenheidsgedichten).**[10] It contains metadata on poems that are among others written to celebrate the marriage of notables and describes poems with the name of the bride and groom, the marriage date, and bibliographic data such as information on the author and publisher.

Description. The use case aims at providing an enriched view on the *creation of poetry* back in the day. One possible outcome is to incite early start on literature by shedding light on the age in which poets started back then motivated by the occasion of, for example, the birth, marriage or death of notables. This can unveil relations between the type of festivities and honored persons and the poets and their ages. By connecting poetries with their corresponding event in SAA, it is likely to enrich biographical information on the honored persons and their social connections, since those datasets contain complementary information that can be used to help disambiguate mentions of people, mostly identified by their names.

Results. The finding of integration links is done using the Lenticular Lens. These links and their respective metadata are loaded into a Stardog 7.8 triple store so that they can be queried, for example, for generic competency questions translated into SPARQL queries in Sect. 4.1. As the metadata gets lengthy, it is not feasible to display it in the current article, but they can be observed by following the links given for the case-study. The use-case results in 4 linksets and 3 lenses, in a total of around 10K links between 14K entities. We present here the means to investigate under which conditions the links are created and combined, if they have been validated and more. Results are available online[11] as well as the corresponding links and metadata[12] which are exported from the tool. Hereby we describe an overview of the matches with statistics and present results of the queries addressing the competency questions.

[9] SAA: https://archief.amsterdam/
This paper uses the RDF version of SAA data published by the Golden Agents project.

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

[11] https://lenticularlens.goldenagents.org/?job_id=90b598f72088ebd0e21446a12e353ffd.

[12] https://github.com/knaw-huc/golden-agents-occasional-poetry/.

Matching Overview. The Occasional Poetry dataset distinguishes *marriage* from *marriage anniversary.* For that and for independent analysis reasons, four linksets are created using the Lenticular Lens. Two that link notables to their notice of marriage (linksets 9 and 11) and two that link them to the baptism of their child (linksets 12 and 13). This then leads to the creation of three lenses: `lens:1` groups couples with the intention of marriage who got married within 6 months (`linkset:11`) with those celebrating at most 50 years of marriage anniversary (`linkset:9`); `lens:2` groups couples celebrating a marriage or a marriage anniversary to those baptising their child (`linkset:12` and `linkset:13` respectively); `lens:3` groups `lens:1` and `lens:2`.

Statistics. The integration modelled in Fig. 4 also highlights the total numbers of resources within each datasets respectively (Poetry:15,650, Marriage:1,197,673 and Baptism:4,889,160) and the number of links found per pair of datasets: 65,978 between Poetry and Marriage and 3,856 between Poetry and Baptism.

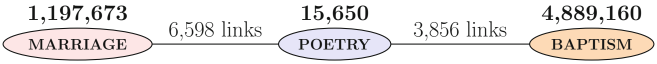


Fig. 4. Occasional poetry matching

CQ 1a & 1b inquire on (a) datasets involved and (b) the restrictions on them for a given link-dataset. Table 1, result of Listing 1.1, shows the metadata of `linkset:9`. It shows two dataset partitions: the City Archives' notice of marriage, in the 1st line, is partitioned simply based on the class roar:Person; the Occasional Poetry, in the next lines, is partitioned based on both the class schema:Person and on a property path that restricts those that are mentioned in a poetry about marriage anniversary (*jarig huwelijk*).

CQ 2a & 2d inquire on the links above a specified threshold that have been rejected given a set of datasets and entity types. Table 2, result of Listing 1.2, shows links with a score above 0.75 yet, listing matched resources (sub/obj), the resource selections from which they originate (subDs/objDs) and the strength resulting from the method applied. This example illustrates that passing the matching method's conditions does not necessarily mean that all resulting links are thereon valid. Often enough, such undesired contextual links need pruning and for that, other techniques such as [8,14] can be applied for a more refined result.

5 Discussion

The development of VoID+ has triggered positive impacts in the humanities community within the Golden Agent project. It has facilitated the development and creation of the Lenticular Lens tool, which is also adopted as link discovery tool in other humanities projects. As the tool relies on VoID+ for its links'

Table 1. Results of QCQ 1a & 1b using Listing 1.1. It shows how datasets are partitioned for a given linkset, namely linkset:8f72088ebd0e21446a12e353ffd-9.

PartLabel	PartType	Restriction	PartitionInFormula	Filter	Value
SAA Notice of Marriage: Person	voidPlus: Class Partition	roar:Person			
OP: Person (marriage anniversary)	voidPlus: Class Partition	schema:Person	AND |- rsc:PropertyPartition-8a..24 |_ rsc:ClassPartition-a6..dd		
OP: Person (marriage anniversary)	voidPlus: Property Partition	1 inv(schema:about) 2 schema:Role 3 inv(schema:about) 4 schema:Book 5 schema:about 6 sem:Event 7 sem:eventType 8 sem:EventType 9 rdfs:label	AND |- rsc:PropertyPartition-8a..24 |_ rsc:ClassPartition-a6..dd	contains	%jarig huwelijk%

Table 2. Results of QCQ-2a & 2d partially displayed using Listing 1.2.
It shows links based on their strengths (> 0.75) and validation flags (rejected).

linkDataset	subDs	sub	objDs	**obj**	**str.**
lens:90..fd-3	OP: Person (marriage anniv.)	stcn: p067702015	SAA Notice of Marriage: Person	saa_deeds: 5e..5f?person=96..c98f..3b	0.84
linkset:90..fd-9	OP: Person (marriage anniv.)	stcn: p067763537	SAA Notice of Marriage: Person	saa_deeds: ab..c7?person=96..1efa..3b	0.88
lens:90..fd-3	OP: Person (marriage)	stcn: p067763537	SAA Baptism: Person	saa_deeds: ab..c7?person=96..1efa..3b	0.88
linkset:90..fd-9	OP: Person (marriage anniv.)	stcn: p069766037	SAA Notice of Marriage: Person	saa_deeds: 87..da?person=96..c29c..3b	0.76
...					

provenance, it ensures that all links discovered by researchers within the project's multiple real-life case-studies are properly and automatically documented. As a result, this allows researchers to now easily generate and consume links with more awareness of the context. However, there is still room for improvements, some of which are pointed out in this section.

VoID+ vs. OWL vs. SPARQL. This work has shown some of the limitations in the state-of-the-art's ontologies when it comes down to the documentation of links and their related processes. In particular, we emphasize here the advancement on complex partitioning offered by the proposed model as the state-of-the-art does not fit the bill. Even though some could argue to use OWL syntax in order to express restrictions such as conjunction, disjunction or property domain [5], we do not believe it is suitable for the description of partitions. That is because partitions are meant to restrict/select subsets of triples in a database, while OWL restrictions are meant to be applied over instances (individuals), eventually producing new triples in case those instances fit the restriction (e.g. class instantiation). Instead, the SPARQL-construct syntax allows exactly for selecting/describing a set of triples of interest based on a pattern (restrictions). In this scenario, VoID+ can be adapted to include the SPARQL-construct syntax.

Vocabulary and Granularity. Question 2.b poses an interesting knowledge representation challenge for link annotation: how much detail or what level of granularity is needed? As observed by [6], more than one metric is often used for resource description comparison in the link discovery process. In such scenario, in order to know which links result from the application of a particular method, one would need to annotate each link with the particular metric by which it is discovered. A rather drastic alternative as we see would be for linksets to be restricted to the use of a single method. However this solution imposes an efficiency issue since linksets simultaneously applying several methods allow for optimization in terms of time and space complexity.

6 Conclusion

Discovering, accessing and understanding the structure and schema of interlinked datasets are areas of metadata covered by VoID vocabulary among others. However, when it comes to (i) reproducing, (ii) understanding or (iii) assessing to reliably reuse datasets of triples interlinking other data, the state-of-the-art ontologies are presented with challenging issues as the concepts and semantics offered do not stretch to concepts that cover the aforementioned needs.

Based on important insights exhibited by competency questions and limitations of existing works ranging from generality, partition ambiguity, semantic restriction and limited concepts, we propose VoID+, an extension for VoID that enables creators, providers and users to document links with sufficient information to reliably support its reuse for context-based data integration.

For testing VoID+, we (1) illustrate generic queries that answer competency questions, (2) present the Lenticular Lens tool for link discovery in-use by projects like Golden Agents and Clariah, and which implements the proposed representation; and (3) discuss a use case of importance in the humanities through the Golden Agents project. With this real-life use case, we successfully show the importance of expanding the state-of-the-art vocabulary to main concepts related to the processes surrounding links in general. We also show how provenance can be extracted using SPARQL as illustrated in Sect. 4.1.

Overall, with the help of researchers within the humanities community and the Lenticular Lens, this paper shows (a) the importance of a full blown and more mature vocabulary for annotating links and their processes, (b) the need to help links creators to automate the annotation of discovered links. Put differently, the need for matching methods to self-document the links creation. Another issue shortly addressed in this work is (c) the usefulness of a flexible and generic tool for managing link-related issues such as the access to off-the-shelf matching algorithms; the generation, manipulation and validation of links.

Future Work. Enrich VoID+ with vocabularies such as PROV-O and SPARQL-construct. Investigate how to combine efforts with similar approaches such as SILK to improve the Lenticular Lens tool. Investigate the nature of computed scores and how to properly combine and manipulate them. For example, understanding during the computation of the final link-score what it means when SILK applies a weight per method to contextually distinguish the ones that are more relevant from the rest.

References

1. Albertoni, R., Pérez, A.G.: Assessing linkset quality for complementing third-party datasets. In: Proceedings of Joint EDBT/ICDT 2013 Workshops, pp. 52–59. ACM, New York (2013). https://doi.org/10.1145/2457317.2457327
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets. In: Bizer, C., Heath, T., Berners-Lee, T., Idehen, K. (eds.) Proceedings of WWW 2009 Workshop on Linked Data on the Web, LDOW, vol. 538, pp. 10. CEUR-WS.org, Madrid (2009)
3. Beek, W., Raad, J., Wielemaker, J., van Harmelen, F.: sameAs.cc: the closure of 500M `owl:sameAs` statements. In: Gangemi, A., et al. (eds.) ESWC 2018. LNCS, vol. 10843, pp. 65–80. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93417-4_5
4. Böhm, C., Lorey, J., Naumann, F.: Creating void descriptions for web-scale data. J. Web Semant. **9**(3), 339–345 (2011)
5. Mendes de Farias, T., Stockinger, K., Dessimoz, C.: VoIDext: vocabulary and patterns for enhancing interoperable datasets with virtual links. In: Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C.A., Meersman, R. (eds.) OTM 2019. LNCS, vol. 11877, pp. 607–625. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33246-4_38
6. Ferrara, A., Nikolov, A., Scharffe, F.: Data linking for the semantic web. Int. J. Semant. Web Inf. Syst. (IJSWIS) **7**(3), 46–76 (2011)
7. Hartig, O.: RDF* and SPARQL*: an alternative approach to annotate statements in RDF. In: Nikitina, N., Song, D., Fokoue, A., Haase, P. (eds.) Proceedings of ISWC, vol. 1963. CEUR-WS.org, Vienna (2017)
8. Idrissou, A., Zamborlini, V., Harmelen, F.V., Latronico, C.: Contextual entity disambiguation in domains with weak identity criteria, pp. 259–262. ACM (9 2019). https://doi.org/10.1145/3360901.3364440
9. Manola, F., Miller, E., McBride, B., et al.: RDF primer. In: W3C Recommendation , vol. 10, no. 1–107, p. 6 (2004)
10. Ngonga Ngomo, A.-C., et al.: LIMES: a framework for link discovery on the semantic web. KI - Künstliche Intelligenz, 413–423 (2021). https://doi.org/10.1007/s13218-021-00713-x
11. Nguyen, V., Bodenreider, O., Sheth, A.P.: Don't like RDF reification?: Making statements about statements using singleton property. In: Chung, C., Broder, A.Z., Shim, K., Suel, T. (eds.) 23rd International World Wide Web Conference, pp. 759–770. ACM, Seoul (2014). https://doi.org/10.1145/2566486.2567973

12. Omitola, T., Zuo, L., Gutteridge, C., Millard, I.C., Glaser, H., Gibbins, N., Shadbolt, N.: Tracing the provenance of linked data using void. In: Proceedings of International Conference on Web Intelligence, Mining and Semantics, WIMS 2011. Association for Computing Machinery, New York (2011). https://doi.org/10.1145/1988688.1988709
13. Orlandi, F., Graux, D., O'Sullivan, D.: Benchmarking RDF metadata representations: Reification, singleton property and RDF. In: 15th IEEE ICSC, CA, USA. pp. 233–240. IEEE (2021). https://doi.org/10.1109/ICSC50631.2021.00049
14. Raad, J., Beek, W., van Harmelen, F., Pernelle, N., Saïs, F.: Detecting erroneous identity links on the web using network metrics. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11136, pp. 391–407. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00671-6_23
15. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., et al. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04930-9_41