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

Al Idrissou[1], Veruska Zamborlini[2,3], and Tobias Kuhn[3]

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

Abstract. This document presents the competency questions, the respective queries designed to answer them and the tables illustrating their application in a particular case study.

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 reuse, 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-type(s), 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 Lens of interest: **(a)** What operators are used to generate the Lens? **(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 and/or Prov-O[3]. However, the level of details required to properly address each question is not achievable with current approaches.

2 Queries for addressing the Competency Questions

Hereby we present some SPARQL queries that allow for answering some of the competency questions posed in Section 1. Listing 1.1 introduces the namespaces required for running the queries in this subsection.

```
BASE             <https://lenticularlens.org/>
PREFIX       rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX      rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX      void: <http://rdfs.org/ns/void#>
PREFIX   dcterms: <http://purl.org/dc/terms/>
PREFIX  voidPlus: <voidPlus/>
PREFIX  resource: <voidPlus/resource/>
PREFIX   linkset: <voidPlus/resource/linkset#>
PREFIX      lens: <voidPlus/resource/lens#>
PREFIX      roar: <https://data.goldenagents.org/ontology/roar/>
PREFIX    schema: <http://schema.org/>
```

Listing 1.1: List of namespaces for running a query

QCQ 1.a & 1.b Given a linkset of interest, the query presented in Listing 1.2 addresses part of question 1 by (a) retrieving interlinked datasets and (b) their explicit partitions by exhibiting selected class, property and/or languages and how they are combined in a formula.

```
SELECT DISTINCT ?PartitionLabel ?PartitionType
?Restriction ?PartitionInFormula ?filterFunction ?filterValue {
  <--- SPECIFY THE GIVEN 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 .
```

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

```
      _: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 1.2: Interlinked datasets & filters (1.a and b)

QCQ 1.c The query in Listing 1.3 answers **question 1.c** as it retrieves the set of methods, their respective supporting metrics and a formula exhibiting the way in which the algorithms are combined for the final discovery of the links belonging to a given linkset.

```
Select ?method ?algorithm ?threshold ?metricRange ?linksetFormula {
  <--- SPECIFY THE GIVEN LINKSET --->
    voidPlus:hasFormulation         ?linksetFormulation .
  ?linksetFormulation
    voidPlus:hasItem                ?method ;
    voidPlus:hasFormulaTree         ?linksetFormula .
  ?method
    voidPlus:hasAlgorithm           ?algorithm .
  OPTIONAL{?method voidPlus:similarityThreshold ?threshold.}
  OPTIONAL{?method voidPlus:similarityThresholdRange ?metricRange.}
  ?algorithm  dcterms:description  ?AlgDescription . }
```

Listing 1.3: Methods for link discovery (1.c)

QCQ 1.d Listing 1.4 complements Listing 1.3 as it addresses **question 1.d** by detailing for a particular method of a given linkset the discriminating properties selected and how they are combined.

```
SELECT DISTINCT
?algorithm ?dsLabel ?propFormulation ?propPartition
(GROUP_CONCAT(DISTINCT ?selProp; SEPARATOR=" \n ") AS ?properties)
{
  <---SPECIFY THE VOIDPLUS METHOD RESOURCE--->
    voidPlus:hasAlgorithm                 ?algorithm ;
    voidPlus:hasObjResourceSelection |
    voidPlus:hasSubjResourceSelection     ?propertySelection.

  ?propertySelection
    voidPlus:hasFormulation               ?formulation .

  ?formulation
    voidPlus:hasItem                      ?propPartition ;
    voidPlus:hasFormulaTree               ?propFormulation .

  OPTIONAL{?propPartition
    voidPlus:subsetOf+/rdfs:label         ?dsLabel . }

  { ?propPartition  void:property  ?selProp .
    FILTER (!isBlank(?selProp)) }
  UNION {
    SELECT ?propPartition
    (GROUP_CONCAT(DISTINCT ?propidx; SEPARATOR=" \n ")
    AS ?selProp)
    WHERE {
      ?propPartition void:property _:pseq .
```

```
      _:pseq
        a      rdfs:Sequence ;
        ?seq   ?prop .
      FILTER (?seq != rdf:type )
      BIND(CONCAT(strafter(str(?seq),"_"), " " ,str(?prop))
      as ?propidx)
    } GROUP BY ?propPartition }
} GROUP BY ?algorithm ?propertySelection ?propFormulation ?dsLabel
         ?propPartition
```

Listing 1.4: Algorithm & settings (1.d)

QCQ 2.a & 2.d Listing 1.5 addresses **question 2** (a) and (d) by retrieving links above a preset threshold of 0.75 given a set of data and entity-types of interest that were rejected by a user. Item (c) can be similarly addressed by selecting the ones accepted instead. However, at present, question 2 can not be answered in its entirety as the ability to properly answer item 2.b particularly challenges the proposed representation.

```
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 1.5: Links rejected yet above a preset threshold of 0.75 given a set of data and entity-types of interest. (2.a and b)

QCQ 3.a & b Listing 1.6 illustrates the retrieval of graphs of links in which a metric of interest is reported to have been used over a given set of properties.

```
SELECT ?linkDataset ?algoLabels ?givenProperty  {
  {
    SELECT DISTINCT ?linkDataset
    (GROUP_CONCAT(DISTINCT ?algoLabel; SEPARATOR=" | ")
    AS ?linkDataset)
    {
      ?algoLinkset
        voidPlus:hasOperand* / voidPlus:hasFormulation /
        voidPlus:hasItem / voidPlus:hasAlgorithm /
        rdfs:label                    ?algoLabel.
    } GROUP BY ?linkDataset
  }
  # a) particular algorithm for link discovery
  FILTER regex(?algoLabels, ".*leven.*", "i")

  # b) set of discriminating properties
  VALUES ?givenProperty { <--- SPECIFY THE PROPERTIES ---> }

  ?linkDataset
    voidPlus:hasOperand* / voidPlus:hasFormulation /
    voidPlus:hasItem                  ?method .

  ?method
    ( voidPlus:hasSubjResourceSelection |
      voidPlus:hasObjResourceSelection ) /
    voidPlus:hasFormulation / voidPlus:hasItem  ?partitionItem  .

  { ?partitionItem        void:property     ?givenProperty }
  UNION { ?partitionItem  void:property     _:pseq .
          _:pseq          ?seq              ?givenProperty }.
}
```

Listing 1.6: Linksets for which a particular metric is used

QCQ 4.a & 4.b Understanding the manipulations that lead to the creation of a lens.

```
select DISTINCT ?lensLabel ?lensDesc ?combination
(GROUP_CONCAT(?operator; SEPARATOR=", ") as ?operators)
{
  <--- SPECIFY THE LENS --->
    a                               voidPlus:Lens ;
    rdfs:label                      ?lensLabel ;
    dcterms:description             ?lensDesc ;
    # a) Retrieving the set-like operator used
    voidPlus:hasOperator/rdfs:label   ?operator ;
    # b) Retrieving the logical operands' combination
    voidPlus:hasFormulation /
    voidPlus:hasFormulaTree           ?combination .
} GROUP BY ?lensLabel ?lensDesc ?combination
```

Listing 1.7: Operators and operands of a lens (4.a and b)

QCQ 4.c For a given lens, Listing 1.8 illustrates the retrieval of all linksets used for its creation as well as there respective associated logical expressions.

```
select DISTINCT ?opLinkset ?linksetFormula ?combination
{
  <---SPECIFY THE LENS--->
    a                           voidPlus:Lens ;
    voidPlus:hasFormulation     ?lensFormulation .

  ?lensFormulation
```

```
    voidPlus:hasFormulaTree     ?combination ;
    voidPlus:hasItem            ?opLinkset .

?opLinkset
    voidPlus:hasFormulation /
    voidPlus:hasFormulaTree     ?linksetFormula   }
```

Listing 1.8: Formulations of a lens' operands (4.c)

3 Evaluation

CQ 1.a & 1b inquire on (a) datasets involved and (b) the restrictions on them for a given link-dataset. Table 1, result of Listing 1.2, 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*).

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 1: Results of QCQ 1a & 1b using Listing 1.2. It shows how datasets are partitioned for a given linkset, namely linkset:90b598f72088ebd0e21446a12e353ffd-9.

CQ 1.c inquires on how links are discovered for a given interlinked dataset. Table 2, result of Listing 1.3, displays some of the metadata of `linkset:13`. Each line shows one applied method, whereas the first column shows a formula combining all of them using an 'AND' operator (minimum T-norm). The second column shows the id of the method followed by the chosen algorithm. We see that `Normalized Levenshtein` was applied twice with a threshold of *7.0*, together with a `Time Delta` algorithm.

method	algorithm	threshold	metric	Range	linkset	Formula
resource: Levenshtein-Normalized-c0..76	resource: Levenshtein-Normalized	7.0		]0, 1]		AND [Minimum t-norm (⊤min)] |—resource:Levenshtein-Normalized-c0..76 |—resource:Levenshtein-Normalized-26..9e |__ resource:Time-Delta-92..76
resource: Levenshtein-Normalized-26..9e	resource: Levenshtein-Normalized	7.0		]0, 1]		AND [Minimum t-norm (⊤min)] |—resource:Levenshtein-Normalized-c0..76 |—resource:Levenshtein-Normalized-26..9e |__ resource:Time-Delta-92..76
resource: Time-Delta-92..76	resource: Time-Delta			*N*		AND [Minimum t-norm (⊤min)] |—resource:Levenshtein-Normalized-c0..76 |—resource:Levenshtein-Normalized-26..9e |__ resource:Time-Delta-92..76

Table 2: Results of QCQ 1.c using Listing 1.3.
It shows which and how algorithms are used and combined for a given linkset: linkset:90b598f72088ebd0e21446a12e353ffd-13.

CQ 1.d inquires on which and how properties are compared for a given method. Table 3, result of Listing 1.4, shows in the first column which algorithm is chosen: `Normalized Levenshtein` in this case. The second column shows the partitions from which the resources to be compared will be steamed: 'Occasional Poetry: Person (marriage anniversary)'and 'SAA Baptism: Person'. Both partitions are further partitioned based on properties selected for the matching purpose. The former is restricted by one property path shown in the 5th column of line 1, while the latter is restricted by two paths shown in the 2nd and 3rd lines. Particularly the 3rd column shows that, when more than one property (path) is selected, a disjunction of them is considered for the partition. This method uses an extra feature called list matching, imposing that a resource in the Occasional Poetry matches with another in the Baptism registries if the names of two people that are mentioned in the same Occasional Poetry (path in line 1) match with the names of a father and a mother mentioned in the same Baptism Record (paths in lines 2 and 3).

algorithm	dsLabel	propFormulation	propPart.	properties
resource: Levenshtein-Normalized	OP: Person (marriage anniversary)	OR [Maximum s-norm (⊥max)] |__ resource:PropertyPartition-02..be	resource: Property Partition-02..be	1 inv(rdf:value) 2 sem:Role 3 inv(sem:hasActor) 4 sem:Event 5 sem:hasActor 6 sem:Role 7 rdf:value 8 schema:Person 9 pnv:hasName 10 pnv:PersonName 11 pnv:literalName
	SAA Baptism: Person	OR [Maximum s-norm (⊥max)] |—resource:PropertyPartition-d0..e8 |__ resource:PropertyPartition-35..ce	resource: Property Partition-35..ce	1 roar:participatesIn 2 thes:Doop 3 inv(roar:carriedIn) 4 **thes:Vader** 5 roar:carriedBy 6 roar:Person 7 pnv:hasName 8 pnv:PersonName 9 pnv:literalName
		OR [Maximum s-norm (⊥max)] |—resource:PropertyPartition-d0..e8 |__ resource:PropertyPartition-35..ce	resource: Property Partition-d0..e8	1 roar:participatesIn 2 thes:Doop 3 inv(roar:carriedIn) 4 **thes:Moeder** 5 roar:carriedBy 6 roar:Person 7 pnv:hasName 8 pnv:PersonName 9 pnv:literalName
• • •				

Table 3: Results of QCQ 1d partially displayed using Listing 1.4.
It illustrates the properties used and how they are combined for a given method, particularly resource:Levenshtein-Normalized-26..9e from Table 2.

CQ 2a & 2d inquire on the links above a specified threshold that have been rejected given a set of datasets and entity types. Table 4, result of Listing 1.5, 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 [**?**,**?**] can be applied for a more refined result.

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
lens:90..fd-3	OP: Person (marriage)	stcn: p067702015	SAA Notice of Marriage: Person	saa_deeds: 5e..5f?person=96..c98f..3b	0.84
lens:90..fd-3	OP: Person (marriage)	stcn: p067702015	SAA Baptism: 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
...					

Table 4: Results of QCQ-2a & 2d partially displayed using Listing 1.5.
It shows links based on their strengths (> 0.75) and validation flags (rejected).

CQ 3a & 3b inquire on the sets of links given an algorithm of interest. Table 5, result of Listing 1.6, illustrates the retrieval of four graphs of links in which the algorithm `Levenshtein` is reported to have been used over the property `pnv:literalName`.

linkDataset	algoLabels	givenProperty
linkset:90b598f72088ebd0e21446a12e353ffd-9	Levenshtein normalized | Time Delta	pnv:literalName
linkset:90b598f72088ebd0e21446a12e353ffd-11	Levenshtein normalized | Time Delta	pnv:literalName
linkset:90b598f72088ebd0e21446a12e353ffd-12	Levenshtein normalized | Time Delta	pnv:literalName
linkset:90b598f72088ebd0e21446a12e353ffd-13	Levenshtein normalized | Time Delta	pnv:literalName
lens:90b598f72088ebd0e21446a12e353ffd-1	Levenshtein normalized | Time Delta	pnv:literalName
lens:90b598f72088ebd0e21446a12e353ffd-2	Levenshtein normalized | Time Delta	pnv:literalName
lens:90b598f72088ebd0e21446a12e353ffd-3	Levenshtein normalized | Time Delta	pnv:literalName

Table 5: Results of QCQ 3a & 3b using Listing 1.6.
Set of graphs composed of links found using a given algorithm (`Levenshtein`) over a set of properties (`pnv:literalName`).

CQ 4.a & 4.b inquire on operators and operands of a given lens. Table 6, result of Listing 1.7, shows a lens obtained by the union of linkDatasets. In particular, the 4th column shows that it is the UNION of two other UNIONS of two linksets each, created themselves as separated lenses (according to the legend in the same cell). In practice, it combines the links resultant from the four linksets.

lensLabel	lensDesc	operators	combination
Person: OP (Occasional Poetry) x Baptism & Notice of Marriage	The union of all links	Union (A ∪ B)	UNION using OR [Maximum s-norm (⊥max)] |—UNION (1.a) using OR [Maximum s-norm (⊥max)] | |—linkset:90b598f72088ebd0e21446a12e353ffd-11 | |__ linkset:90b598f72088ebd0e21446a12e353ffd-9 |__ UNION (1.b) using OR [Maximum s-norm (⊥max)] |—linkset:90b598f72088ebd0e21446a12e353ffd-12 |__ linkset:90b598f72088ebd0e21446a12e353ffd-13 Legend: (1a) created as lens:90b598f72088ebd0e21446a12e353ffd-1 (1b) created as lens:90b598f72088ebd0e21446a12e353ffd-2

Table 6: Results of QCQ-4a&b using Listing 1.7.
Fetching the operators used to manipulate linksets / lenses for the creation of lens:90b598f72088ebd0e21446a12e353ffd-3.

CQ 4.c inquires on the origin of the links in a given lens. Table 7, result of Listing 1.8, shows all the linksets ultimately used for the creation of the lens, as well as the combination of methods employed in each of them.

combination	Linkset	linksetFormula
UNION using OR [Maximum s-norm (⊥max)]	linkset:	AND [Minimum t-norm (⊤min)]
|–UNION (1a) using OR [Max. s-norm (⊥max)]	90..fd-9	|–resource:Levenshtein-Normalized-70..bd
| |—linkset:90..fd-11		|–resource:Levenshtein-Normalized-74..c3
| |__ linkset:90..fd-9		|_ resource:Time-Delta-24..c8
|–UNION (1b) using OR [Max. s-norm (⊥max)]	linkset:	AND [Minimum t-norm (⊤min)]
|—linkset:90..fd-12	90..fd-11	|–resource:Levenshtein-Normalized-93..41
|__ linkset:90..fd-13		|–resource:Levenshtein-Normalized-c7..18
		|_ resource:Time-Delta-71..4d
Legend:	linkset:	AND [Minimum t-norm (⊤min)]
(1a) created as lens:90..fd-1	90..fd-12	|–resource:Levenshtein-Normalized-cb..d2
(1b) created as lens:90..fd-2		|–resource:Levenshtein-Normalized-73..e8
		|_ resource:Time-Delta-72..4d
	linkset:	AND [Minimum t-norm (⊤min)]
	90..fd-13	|–resource:Levenshtein-Normalized-c0..76
		|–resource:Levenshtein-Normalized-26..9e
		|_ resource:Time-Delta-92..76

Table 7: Results of QCQ 4.c using Listing 1.8.
It illustrates a comprehensive formulation at the origin of the sets of links involved in the creation of a lens which is in this scenario lens:90b598f72088ebd0e21446a12e353ffd-3.