A Linked Data Model for Data Scopes

Victor de Boer[1][0000−0001−9079−039X], Ivette Bonestroo[1], Marijn Koolen[2][0000−0002−0301−2029], and Rik Hoekstra(2)[0000−0002−6951−8014]

[1] Vrije Universiteit Amsterdam, the Netherlands v.de.boer@vu.nl
[2] Huygens ING, the Netherlands {rik.hoekstra, marijn.koolen}@di.huc.knaw.nl

Abstract. With the rise of data driven methods in the humanities, it becomes necessary to develop reusable and consistent methodological patterns for dealing with the various data manipulation steps. This increases transparency, replicability of the research. Data scopes present a qualitative framework for such methodological steps. In this work we present a Linked Data model to represent and share Data Scopes. The model consists of a central Data scope element, with linked elements for data Selection, Linking, Modeling, Normalisation and Classification. We validate the model by representing the data scope for 24 articles from two domains: Humanities and Social Science.

Keywords: Scholarly data · Linked Data · Data Scope.

1 Introduction

In recent years, digital tools and methods have permeated the humanities domain [4]. With more collections and archives being digitized as well as the growth of 'digital born' data, a Digital Humanities (DH) movement has gained popularity [2]. While digital data and tools can make humanities research more effective and efficient and uncover new types of analyses, as long as the methods used are transparent and reproducible methods. Adhering to principles of FAIR data management [11] will not only increase the reusability of digital data and methods, but also ensure that they can be subjected to the same rigorous criticism of tools and data as is common in humanities research [7].

The paper 'Data scopes for digital history research' introduces the concept of "data scopes" to alleviate a lack of transparency and replicability with regards to the data manipulation steps in historical research. Data scopes are proposed to "characterize the interaction between researchers and their data and the transformation of a cluster of data into a research instrument" [6].

The original Data scopes paper presents a qualitative model including the five data manipulation activities. We here present an open standardised machine readable format for data scopes to further increase transparency and reproducibility by allowing for (semi-)automatic analysis and replication. It also moves the model even more towards the FAIR principles, making the data scopes a method to publish data manipulation steps as findable, accessible, interoperable and reusable. The model we present here is expressed using Linked Data

principles [5], where we use the Resource Description Framework (RDF) to define an ontology defining the concepts and relations for the model based on [6].

Next to presenting the ontology and its design decisions, we also provide an initial validation of the model in Section 4 by manually annotating articles from two research domains, that of (computational) humanities and social science.

2 Related Work

The Nanopublications ontology allows for FAIR and machine-readable representations of scientific claims and assertions [3]. This has created more incentives for researchers to use this standard format which increases the accessibility and interoperability of the information. Related to this is the PROV model that allows for specifying (data) provenance[8]. The combination of Nanopublications and PROV provides a powerful mechanism to express generic scientific statements and their provenance. The model we present here is compatible with these models, yet provides more specific detail towards DH use cases.

SPAR (Semantic Publishing and Referencing) is a comprehensive set of ontologies describing concepts in the scholarly publishing domain [10]. These include the Document Components Ontology (DoCo) that describes different aspects related to the content of scientific and scholarly documents. DoCo consist of three parts: document components, discourse elements and a pattern ontology. This ontology improves interoperability and shareability of academic documents. The model we present here can be used in combination with SPAR and DoCo to not only describe a research document, but also the data manipulation steps taken in the research, and the context in which the conclusions are valid.

3 Design of the Data Scope ontology

The central concept of a "data scope" as introduced by Hoekstra and Koolen [6], describes a view on research data as well as the process that results in this view. The process is inherently iterative and includes modelling decisions, interpretations and transformations on the data. The datascope takes shape through five activities:

- **Selection**: which data and sources are selected? This matches the process of forming a corpus for a specific research question.
- **Modelling**: how are the relevant elements in sources represented? With the increased use of digital tools, these models become more and more explicit.
- **Normalization**: how are surface forms mapped to a normalized form? (e.g. the mapping of person names to a "Firstname, Lastname" form.)
- **Linking**: what explicit internal and external connections are established? This includes actions like deduplication, named entity resolution etc.
- **Classification**: how are objects grouped or categorized? This includes categorization using internal or external schemes or theories.

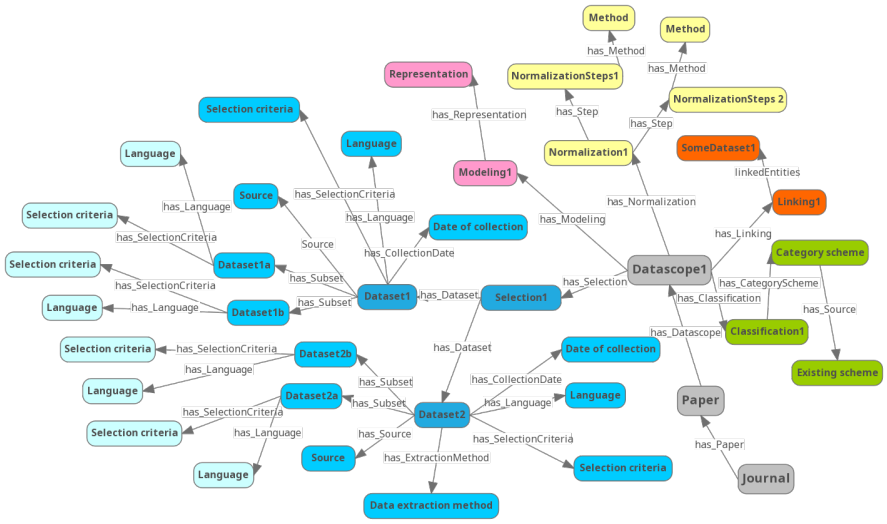

Fig. 1. The data scope model. Boxes show class instances and arrows depict object relations. The colors show the five parts of the model. Classes and namespaces are omitted for brevity.

For our ontology, we used the data scope and five activities as the basis of the model. Further classes and properties are derived from the activities descriptions in [6] of the components to help establish the steps or classes that can be linked to each component. Finally, we selected six research articles of digital humanities and computational social science articles as samples to adapt and adjust the model. These articles come from the same pool as our evaluation data set, which we describe in Section 4.

The resulting ontology contains 14 classes, 14 object properties and 4 datatype properties. Its classes include those for the Data scope itself, the five activities, and subactivities such as `dsont:NormalizationStep`, to define a separate step in the normalization procedure. Classes for research articles allow for associating a data scope to a publication. Figure 1 shows the ontology by means of an abstract example, where class instances are depicted (for example, "Datascope1" is an instance of `dsont:DataScope`). A DataScope instance links to instances of each of the five activities. Each activity allows for further specifications. For example the selection part links to the selection of the datasets. These datasets link for example to the date in which they have been collected etc.

The ontology is expressed in RDFS and is available on github[3]. The data-model uses permanent w3id identifiers for its dereferenceable URIs (namespace `https://w3id.org/datascope/`). The ontology, example data and annotation results (see Section 4) can be queried using SPARQL at `https://semanticweb.`

[3] `https://github.com/biktorrr/datascope_ontology`

cs.vu.nl/test/user/query. The sample query below counts for each data scope the number of datasets for which a normalization step is registered.

```
PREFIX dsont: <https://w3id.org/datascope#>
SELECT ?s ?norm (COUNT(?ds) as ?dscount) WHERE
{ ?s rdf:type dsont:DataScope .
?s dsont:has_Selection ?sel .
?sel dsont:has_Dataset ?ds .
?s dsont:has_Normalization ?norm }
GROUP BY ?s ?norm
```

4 Model validation

We perform an initial validation of the model by manually annotating 24 articles using the model.

We selected 24 articles from two related domains: humanities and social science, focusing on publications that include digital data as part of the methodology. To this end, we selected two research journals that focus specifically on digital methods in these two fields: Digital Scholarship for the Humanities (DSH)[4] and Computational Social Science (CSS)[5]. We selected articles published after 2018, resulting in an initial selection of 124 articles from DSH and 58 articles from CSS. These articles were filtered on the inclusion of a clear data section,which resulted in 71 articles. Of those selected articles, we randomly chosen 15 articles of each journal. 6 of the 30 articles were used in the design phase of the model (Section 3). The remaining 24 articles have been used for the validation.

For the annotation, a set of coding guidelines were established based on the data scopes paper as well as the data scopes ontology description. Two independent coders then each annotated the data sections of the 12 articles using these coding schemes. Each article is mapped to the ontology and expressed as RDFS instances of the classes. If activities or steps are not explicitly identified in the text, they are not represented as RDF triples. This gives us an indication of the coverage of the various classes and properties in current articles. Every step in an article that did not quite fit the model or the concept of data scope was noted in a document. We have only looked at the sections describing the data and did not look at other sections. In some cases, that meant that pre-processing steps were not recorded. In cases where such pre-processing steps actually make changes to the data, these should be considered part of the datascope. This shows that it is not always straightforward to identify the limit of a data scope (this was already identified in [6]).

Figure 2 shows for the main classes how many articles of the two journals have *at least* one element that was identified in the annotation. This includes the five main activities plus the main sub-elements of the Selection class ("Sel-").

Regarding the five main classes, we can see that the selection part is described in all articles, classification in 8, normalization in 4 and linking and modelling are both described in only one article. This matches the prediction by Hoekstra

[4] https://academic.oup.com/dsh

[5] https://www.springer.com/journal/42001

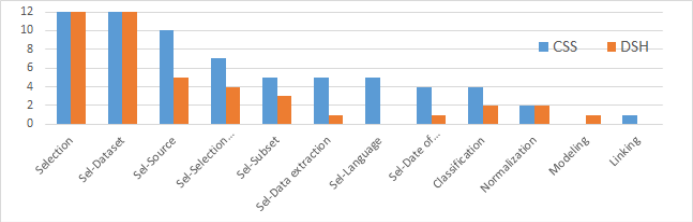

Fig. 2. Results of the annotation using the data scopes ontology for 24 articles.

and Koolen that the Selection is most likely the element that is most often described currently. The figure does not show large differences between the two journals, the main discrepancy being one or two counts in either direction for classification, linking, and modelling. Looking at the Selection elements, we can see that the classes of the first dataset in the selection components are used by all the articles. We here identify some differences: articles from CSS mention all the steps within the selection part of our model more often than articles of DSH. Articles of DSH almost never mention language in the steps of the first dataset.

While each of the classes is used at least once in our 24 articles, many steps are not represented in the current papers. This can be a reflection of a lack of these activities, or of expressing these activities explicitly in the resulting papers.

5 Discussion

In our annotation effort, we have seen that in some cases it is unclear for some data manipulation step to which data scope elements they should be mapped. One example is the article by Badawy and Ferrara[1], where an identification step occurs before the selection process. Here the authors identify which Twitter accounts belong to ISIS sympathizers before selecting the data.This could be mapped to a selection or classification activity. Other such choices occur between for example normalization and classification. For replicability, clear guidelines on how the different elements of the model are used should be provided.

Currently, the ontology is quite high-level with a limited amount of classes and properties. It can be further specified towards concrete cases. For example, the research by Mantzaris et al. [9] uses multiple classification schemes because of changes in vote distribution for Eurovision throughout the years. This complexity does not fit our model yet but classes for this could be added.

Further possible extensions identified in the articles include classes about the data timeframe, specific information about dataset contents, filter options for data extraction, definitions of classifications, multiple classification schemes and interconnection between components.

Our model is compatible with models such as PROV and DoCo. DoCo information can be linked through the elements component which includes data and method in which the data scope is described. It is future research to combine the

data scopes ontology with that of NanoPublications, that can be used to publish data scopes, further increasing the findability and verifiability of the research.

We here only provide an initial validation of the data model for a limited number of articles. These articles are selected to have data manipulation steps. We expect that in non-digital humanities articles, we will see hardly any occurrences of explicit data manipulation steps and therefore, the statistics presented before cannot be extrapolated beyond this selection. However, with the growing interest in digital tool criticism [7] and comprehensive virtual research environments, we expect that more such information will be made available. Current research consists of integrating the model in such a virtual research environment.[6] With more and more scholars using such environments and other digital tools and data, the data model we present in this paper represents a considerable step towards transparent and reproducible digital humanities and social science.

Acknowledgements. The research for this article was made possible by the CLARIAH-PLUS project financed by NWO (`http://www.clariah.nl`). The authors would like to thank Gard Ruurd for his contributions to the research.

References

1. Badawy, A., Ferrara, E.: The rise of jihadist propaganda on social networks. Journal of Computational Social Science **1**(2), 453–470 (2018)
2. Berry, D.M.: Introduction: Understanding the digital humanities. In: Understanding digital humanities, pp. 1–20. Springer (2012)
3. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use **30**(1-2), 51–56 (2010)
4. Haigh, T.: We have never been digital. Communications of the ACM **57**(9), 24–28 (2014)
5. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology **1**(1), 1–136 (2011)
6. Hoekstra, R., Koolen, M.: Data scopes for digital history research. Historical Methods: A Journal of Quantitative and Interdisciplinary History **52**(2), 79–94 (2019)
7. Koolen, M., Van Gorp, J., Van Ossenbruggen, J.: Toward a model for digital tool criticism: Reflection as integrative practice. Digital Scholarship in the Humanities **34**(2), 368–385 (2019)
8. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: Prov-o: The prov ontology. W3C recommendation **30** (2013)
9. Mantzaris, A.V., Rein, S.R., Hopkins, A.D.: Preference and neglect amongst countries in the eurovision song contest. Journal of Computational Social Science **1**(2), 377–390 (2018)
10. Peroni, S., Shotton, D.: The spar ontologies. In: International Semantic Web Conference. pp. 119–136. Springer (2018)
11. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(1), 1–9 (2016)

[6] http://mediasuite.clariah.nl