Swirling and Shackling Literary Graphs

Susan Brown[1], Deborah Stacey[2], and Alliyya Mo[2]

[1]School of English and Theatre Studies, University of Guelph
[2]School of Computer Science, University of Guelph
[3]LINCS Project, University of Guelph

1 Introduction

The Linked Infrastructure for Networked Cultural Scholarship (LINCS) infrastructure project is converting cultural datasets in a range of forms to Linked Open Data with a view to improving findability, interoperability, and reusability (LINCS (2021)). In converting or extracting granular linked data from researcher source data content alongside object metadata, it differs from large-scale cultural heritage and humanities projects focused on linked open metadata for large collections or aggregations of research objects. The linked data produced by LINCS is granular but uneven in its level of detail, given its provenance in individual scholarly projects and the gaps in historical records. The project thus faces the challenge of creating data robust enough to support not only querying and browsing but also reasoning across graphs generated from different interrelated but uneven datasets.

We aim to test the potential of the Semantic Web Rule Language (SWRL) rules and Shapes Constraint Language (SHACL) to generate new relationships from existing ones. Medical researchers are employing SWRL and SHACL to support reasoning with complex patterns in health data (Lezcano et al. (2011); Pareti et al. (2019); Somodevilla et al. (2021)). Similar methods can be applied to cultural data, for instance to move beyond direct links between individual people to generate speculative "webs of textuality" or "webs of influence" (second-degree or friend-of-a-friend relationships, for instance) that can then be evaluated by constraints in SHACL to see which parts of these webs are most likely or possible. Because cultural datasets are far from comprehensive, this strategy, may help to "fill in the gaps" in the historical record. This speculative mode of building on existing data aims to prompt new inquiry into underrepresented activities and figures and suggest ways in which silences in historical data might be tackled computationally. In addition, because most LINCS data is being generated from datasets that were not created with RDF representation in mind, SHACL and SWRL present the potential for improving data quality by weeding out logically impossible statements that are being generated via scripts, including from the XML-encoded prose of the literary historical textbase Orlando: Women's Writing in the British Isles from the Beginnings to the Present (Brown et al. (2019a, 2021)).

2 The Power of Rules: SWRL

Semantic Web Rule Language (Horrocks et al. (2004)) enables Horn-like logic rules to be combined with an OWL knowledge base. All of the logic rules are expressed in terms of OWL concepts such as classes, properties, and individuals. Rules are of the form of an implication between an antecedent (body) and consequent (head) and can be read as: "whenever the conditions specified in the antecedent hold, then the conditions specified in the consequent must also hold". A simple example demonstrates how knowing parent and sibling relationships can be used to produce the relationship "aunt". The rule: $hasParent(?x1, ?x2) \wedge hasSister(?x2, ?x3) \Rightarrow hasAunt(?x1, ?x3)$ says that if $x1$ has a parent $x2$ and $x2$ has a sister $x3$, then $x3$ is the aunt of $x1$. So, if "Edith hasParent Emma" and "Emma hasSister Katharine" then "Edith hasAunt Katharine."

SWRL can play an important part in using logic to "create" statements that are implied by statements already in the knowledge base. It is not always possible to explicitly define all the relationships that exist in a knowledge base, but rules can "fill in the blanks" and ideally should be defined at the same time as the ontology is designed or when an ontology is adopted for use.

2.1 Looking for Connections

History is always partial and incomplete, women's literary history even more so. We are experimenting with compensating for the unevenness of the data to use SHACL/SWRL to hypothesize connections that may not be explicit.

Did writer A and B have a professional relationship? If sharing a publisher is indicative of a "weak" professional relationship (i.e. possible but without direct evidence) then a set of SWRL rules that checks that their years of working with the same publisher overlap could create a relationship such as:

```
IF
    writerA is_published_by publisherA AND
    writerA published_starting_year yearX1 AND
    writerA published_ending_year yearY1 AND
    writerB is_published_by publisherA AND
    writerB published_starting_year yearX2 AND
    writerB published_ending_year yearY2 AND
    ((yearX2 >= yearX1 AND yearX2 <= yearY1) OR
     (yearY2 >= yearX1 AND yearY2 <= yearY1))
THEN
    writerA has_weak_prof_relat_with writerB
```

The combination of many collections of such statements could profit from rule sets that establish relationships that would not be present in the collections themselves.

Such methods may also help researchers to evaluate existing hypotheses. For instance, the writings of Dante Gabriel Rossetti (Rossetti (1870)) include the poem "Jenny" published in April 1870, two months after Augusta Webster's "A Castaway" was published (Webster (1878)). Both are dramatic monologues spoken by a prostitute, approached very differently; Webster's reads like a rebuttal to Rossetti. Moreover, Dante's sister Christina Rossetti's posthumously published critique (Rossetti (1896) of vampiric representation is echoed by Webster. A hypothesis of influence in both cases might, however, be flagged by SHACL rules as doubtful, either through publi-

cation dates or because no direct relationship between Dante Rossetti and Webster is asserted in the Orlando data. However, all three authors published with MacMillan in overlapping periods, allowing SWRL rules to posit weak professional relationships amongst them. In addition, the data records Webster's social and political relationships with Christina and her other brother, William Michael. The combination of these multiple relationships, some established and some hypothesized, supports the presumption that Webster might well have had access to both texts in manuscript form even though direct evidence is lacking.

Most writers in the Orlando datasets are connected either directly or at one degree of separation via another person, place, organization, or text. Some connections are trivial in the case of well-connected figures such as Charles Dickens–while others are profound, such as Dickens' relationship to Elizabeth Gaskell. Going forward, we aim to use SWRL to establish which types of relationships, either solo or in combination, can be leveraged to hypothesize meaningful connections between individuals or to filter them in useful ways. We hope such methods can help sift the dense networks of relationships surrounding publishers, journals, and individuals represented in the data, complementing social network analysis methods and allowing us to get at meaningful, but not necessarily prominent, clusters of relationships.

3 SHACL: Using constraints to automate validation

SHACL, or Shapes Constraint Language, is a language for validating RDF graphs against a set of conditions. The SHACL "constraints" are expressed as RDF graphs known as "shape graphs" and the RDF graphs that are to be validated are called "data graphs". Shape graphs are used to check that data graphs satisfy a particular set of conditions (Knublauch and Kontokostas (2017)).

The use of constraints to describe RDF data has many uses, since validation of data is at the core of many LOD activities. One example of the use of constraints for validation of RDF graphs is the use of SHACL to determine if reconciliation has been done correctly. If, for example, some tool in our pipeline has identified a particular person as the mother of another person, and then reconciled this with a known person (someone in Wikidata or DBpedia, for example), then we can create a description of "mother" that will help to determine if our reconciliation was reasonable. A mother description could include such constraints as age (if both birthdates are known then we can see if it is reasonable for the person to be a parent), gender, place of residence (if the potential father only ever resided in England and the potential child spent their entire life in Australia, then the relationship is possible but not likely).

3.1 Impossible relationships

In an earlier version of the Orlando graph (Brown et al. (2019b)), for instance, the Python scripts used to create RDF produce the assertion that Margaret Atwood is the son of Susanna Moodie, due to the inclusion of Atwood's name within the XML tag

<FAMILY > <MEMBER RELATION=SON >

The location information (both inhabited Ontario) would be insufficient to rule out the relationship, but the impossibility of someone born in 1803 being the mother of someone born in 1939, as well as the unlikelihood of a mother-son relationship occurring between two individuals with the assigned gender of "woman", could rule

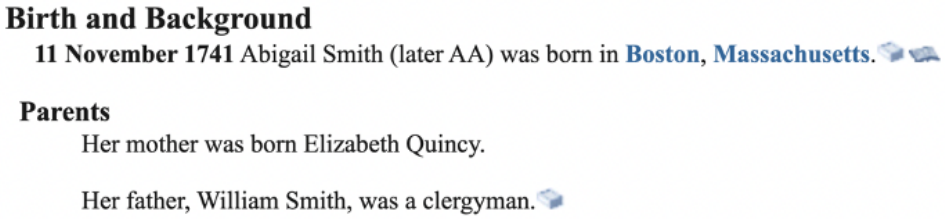

Figure 1: Excerpt from Orlando source data for Abigail Adams

out or flag the creation of such relationships.

Similarly, Shakespeare is mentioned more than 1100 times in the Orlando dataset; he is both a noteworthy literary figure and a major influence on women writers. His ubiquity therefore makes him a prime candidate for data extraction that result in false assertions, since, like Atwood, he sneaks into XML tags that reference other people: for example, in 8 instances of the tag designating intimate relationships, only one directly involves Shakespeare. Rules that require that the lifetimes of those involved in intimate relationships must overlap would be useful for other types of relationships as well.

3.2 Weeding out automated reconciliation errors

LINCS is adapting NLP technologies based on the Diffbot knowledge graph for reconciliation and relationship extraction, but the Diffbot tools lean towards contemporary data and use cases related to commercial applications; the results need to be refined when being used with historical data. We are evaluating the results in part by seeing how well Diffbot's API (Diffbot (2021)) identifies entities and relationships in untagged versions of the XML text from which we are generating RDF, such as a snippet from the Orlando profile for Abigail Adams (Figure 1).

The profile for Adams (1744-1818), First Lady of the United States, asserts, "Her father, William Smith, was a clergyman." Diffbot's reconciliation (Figure 2) finds a William Smith in DBpedia who was born in 1819 and died in 1892 (DBpedia (2021)).

This name is a challenge for reconciliation: in Wikidata, multiple William Smiths are retrieved including Q313533 (William Smith, 1769-1839) and Q96076033 (William Smith, 1707-1783). The William Smith retrieved by Diffbot from DBpedia is described as, "William Smith (1819–1892) was a Catholic clergyman from Scotland. He served as the Archbishop of the Archdiocese of St. Andrews and Edinburgh." Since Adams was born in 1744, DBpedia's William Smith could not be her father since he was not born until 1819. Thus, he fails the hard constraint of a child being born within the lifespan of the parent (for fathers, lifespan + 9 months). By hard constraint we mean that this constraint cannot be violated. We can also have "softer" constraints such as parents and children sharing the same ethnicity or nationality. And while the text snippet about Abigail Adams does identify her father as a "clergyman", the soft constraint that Roman Catholic bishops are not supposed to produce offspring should only trigger some kind of warning, since this constraint can be violated while the parental relationship holds true!

SHACL can validate the "completeness" of certain kinds of information. A combination of SHACL and SWRL, or SHACL and SPARQL can be constructed to validate that children are born within the lifespan of their parents. This type of validation

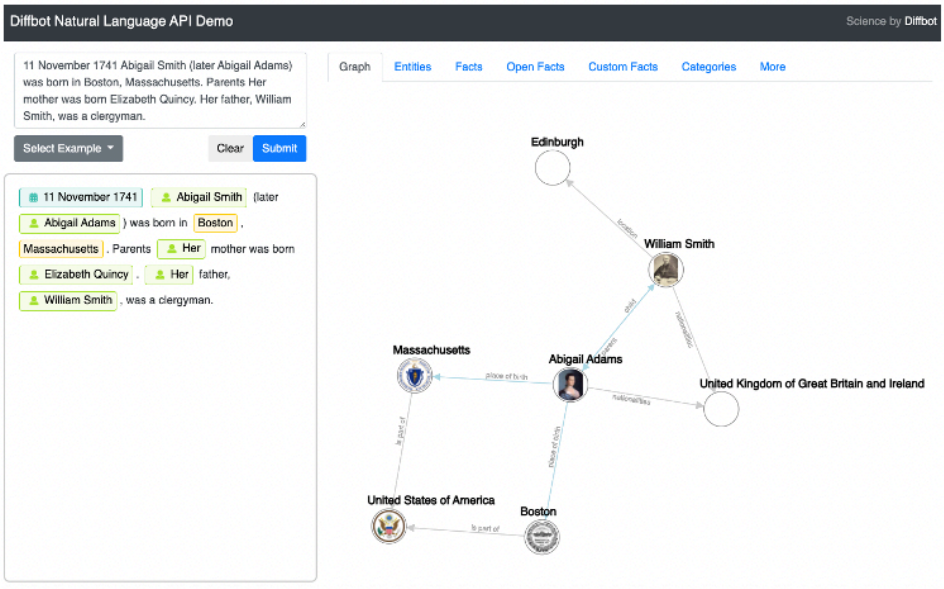

Figure 2: Visualization of reconciliation and relationship results using the Diffbot API Demo

may be used after other SHACL shapes have checked that all persons identified as parents actually have dates of birth and death. This type of check generates warnings that these individuals will be difficult to use in other validation exercises.

SHACL can also be used to identify inconsistencies that require human vetting. This has potential to automate validation in the data processing pipeline and reduce tedium for researchers. Hard constraints (e.g. a father cannot be born after his daughter) can be detected and possibly handled automatically by an editing process while soft constraints (e.g. parents and children usually share the same ethnicity) can be flagged for vetting by human experts. Flagging for human attention will be very helpful in LINCS, where outliers may be of particular interest for cultural analysis, so we do not want to over-regularize the data.

The paper aims to demonstrate that SWRL and SHACL provide useful complements to SPARQL for the analysis, data review, and quality assurance of cultural datasets. It provides a preliminary sense of the opportunities and limitations of the approach, based on initial experimentation using the CWRC ontologies (CWRC (2021)). Following this initial investigation, next steps for this work include applying these techniques to the same data in LINCS format, which uses CIDOC CRM, and comparing with related work with cultural data (Faraj and Micsik (2021); Lasolle et al. (2021)).

References

Brown, S. et al. (2019a). Orlando Archived Dataset on Scholars Portal, V1. https://sparql.cwrc.ca.

Brown, S. et al. (2019b). The Orlando British Women's Writing Dataset Release 1: Biography and Bibliography. https://doi.org/10.5683/SP2/EOB9S6.

Brown, S. et al. (2021). Women's Writing in the British Isles from the Beginnings to the Present. Cambridge University Press.

CWRC (2021). The CWRC Ontology Specification 0.99.86. https://sparql.cwrc.ca/ontologies/cwrc.html.

DBpedia (2021). About: William Smith (bishop). https://dbpedia.org/page/William_Smith_(bishop).

Diffbot (2021). Diffbot API Demo. https://demo.nl.diffbot.com/.

Faraj, G. and A. Micsik (2021). Representing and Validating Cultural Heritage Knowledge Graphs in CIDOC-CRM Ontology. Future Internet 13(13). https://www.mdpi.com/1999-5903/13/11/277/htm.

Horrocks, I. et al. (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML. https://www.w3.org/Submission/SWRL/.

Knublauch, H. and D. Kontokostas (2017). Shapes Constraint Language (SHACL). https://www.w3.org/TR/shacl/.

Lasolle, N. et al. (2021). A Semantic Web Navigation Tool for Exploring the Henri Poincaré Correspondence Corpus. In Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage. Hal Open Science. https://hal.univ-lorraine.fr/hal-03406713.

Lezcano, L. et al. (2011). Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. Journal of Biomedical Informatics 44(2). https://doi.org/10.1016/j.jbi.2010.11.005.

LINCS (2021). Linked Infrastructure for Networked Cultural Scholarship. https://lincsproject.ca/.

Pareti, P., G. Konstantinidis, T. J. Norman, and M. Şensoy (2019). SHACL Constraints with Inference Rules. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon (Eds.), The Semantic Web – ISWC 2019, Cham, pp. 539–557. Springer International Publishing.

Rossetti, C. (1896). New Poems, Hitherto Unpublished or Uncollected. Macmillan.

Rossetti, D. G. (1870). Poems. Macmillan.

Somodevilla, M. J., I. Mena, I. H. Pineda, and M. C. P. d. Celis (2021). Deducting Lifestyle Patterns by Ontologies' SWRL Rules,. In 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), pp. 9–13. Hal Open Science.

Webster, A. (1878). Portraits. Macmillan.