COST STSM Reference Number: 201010
COST Action: IC0801 (WG1)
STSM Topic: Incorporating Provenance-related Annotations for Linked Data Reasoning
Applicant: Aidan Hogan (DERI Galway)
Host Institution: Universitá di Napoli (Federico II)
Location: Napoli, Italy
Duration: 2010-07-05 – 2010-07-11

Purpose of the meeting
The goal of this STSM was to meet and work with Prof. Piero Bonatti and Dr. Luigi Sauro – Piero had previously visited out institute in Galway, and during this visit we had made an initial agreement to collaborate on using an annotated Logic Programs framework to integrate ranking values and notions of trust and provenance into our current Linked Data reasoning approach. The purpose of my visit was to help make the plans for our work in this area more concrete, to get a better understanding of the theoretical aspects of the work, to give Piero and Luigi a better indication as to the requirements of the use-case, and to help derive a research plan for further pursuance of this joint work.

Overview
Together with Dr. Axel Polleres at DERI Galway (National University of Ireland, Galway), we have been working for a number of years on scalable rule-based reasoning, in particular performing incomplete materialisation over large-scale openly-published Linked Data corpora: reasoning over such data implies unique challenges, including (i) scalability where we investigate reasoning over data in the order of a billion triples; (ii) tolerance to noisy and conflicting Web data, where publishers are prone to making errors and inconsistent statements and where incorporating the source of information is crucial to prevent nonsensical and potentially explosive inferences. As such, we have developed and published the details of a mature approach: SAOR (Scalable Authoritative OWL Reasoner [1]) which relies on a separation of terminological knowledge and distribution to efficiently reason over datasets in the order of a billion triples, and which includes numerous adaptations to be tolerant to noisy Web data, where in particular we introduced the notion of authoritative reasoning which only allows certain T-Box axioms to be considered for reasoning if they appear in "authoritative" documents. In particular, this work has been inspired by (and is actively used by) the SWSE system [2] for query-answering over Linked Data corpora.

However, our approach thus far has been very engineering-oriented, discussing specific problem-driven optimisations and algorithms for solving the given challenges. Also, we wished to extend the approach further: in the data, there exist many forms of inconsistencies which we currently ignore [3]. Although we can detect these inconsistencies, we want to go further and try to resolve/repair these data issues: colleagues at DERI have been working on ranking algorithms for Linked Data sources and we wished to investigate incorporation of such an algorithm for deciding for which of the inconsistent views the dataset provides the most evidence.

In this respect, Piero and Luigi's theoretical expertise on annotated Logic Programs is pertinent. Using an annotation framework, we can – in a principled and well-understood manner – provide the SAOR reasoner with additional meta-data which encodes information relating to the source of (Web document providing) artefacts of data, and also propagate information relating to results of our ranking procedure. Using annotations, not only can we give a better theoretical grounding for our current authoritative algorithm, but also extend our functionality to hopefully resolve inconsistencies in the data, increasing the quality of the resulting inferences for our use-case. Indeed, using annotations further allows for future extensibility in the same manner.

As such, the work is interesting wrt. two main aspects: (i) we can generalise how SAOR incorporates information about provenance into the reasoning algorithm and look at further extensions useful for the

Linked Data reasoning use-case; (ii) we can demonstrate a new use-case for annotated Logic Programs: incorporating provenance-related metadata with reasoning over Web data.

Working diary

• *Monday, 5th July, 2010:* Axel Polleres and I arrived in Naples. We had dinner with Piero and some initial discussions about the objectives of the week.

• *Tuesday, 6th July, 2010:* Axel and I spent our first day at the Universitá di Napoli. Since this was Axel's only day, Axel, Piero, Luigi and I had some high-level discussions on the requirements of the work. Piero described on a high-level the annotated logic program framework, and in particular we discussed aggregation functions for including ranking values in the reasoning procedure. We also briefly discussed on a high level how the ranks could be used to filter or block certain inferencing paths during reasoning, had a brief review of related work, and reviewed possible venues for publishing our future work.

• *Wednesday, 4th August, 2004:* Piero, Luigi and I discussed the use of contra-positives in the inferencing procedure: using constraints to infer negative literals. This approach initially seemed promising, and I believed could be applied at large scale. I had detailed discussion with Luigi about the current authoritative reasoning algorithm, where we discussed some possible generalisations of the approach. We also discussed the formalisation of the existing authoritative reasoning prototype in terms of annotations which Luigi had been working on. I also setup a small test-corpus for running some initial experiments: I extracted ~8m triples from a larger Linked Data corpus and applied the ranking procedure to derive an initial test-bed for exploratory experiments.

• *Thursday, 5th August 2004:* Piero, Luigi and I again discussed the use of rank-value annotations to block certain inferences during reasoning: Piero expressed concerns about computational complexity and the lack of an efficient implementation and thus we decided to abandon this possible optimisation. We also discussed the use of contra-positives, where Piero was concerned that only using constraint rules to infer contra-positives seemed arbitrary: subsequently, Luigi and I reviewed the OWL 2 RL rules (which we intend to apply) and found that contra-positives were not efficiently compatible with `rdfs:domain`/`rdfs:range` rules. Thus, we decided instead to abandon contra-positives and perform inconsistency-detection during post-processing of the inferred data.

•*Friday, 6th August, 2004:* I had been working on applying some of the methods discussed over the 8m triple test-dataset. I had derived some initial results showing examples of the inconsistencies that occur, which included the rank values for the involved triples. Many inconsistencies are created by invalid data-types, with the rest being memberships of the intersection of classes defined to be disjoint. We saw that inconsistencies occur not only in low-ranked documents, but evenly across a spectrum of ranks. For memberships of disjoint classes, we saw examples where the lowest ranked membership was clearly incorrect (e.g., one example we encountered was where the claim "W3C is an organisation" was ranked higher than "W3C is a person"). However, we also saw examples where there was no clear right or wrong side to the inconsistency.

Main Results

- We derived a mature set of requirements for the annotation framework – including annotations for blacklisting, authoritativeness and ranking metrics – which Piero has now formalised.
- We discussed the integration of the authoritative reasoning algorithm in the annotation framework and possible means of generalising the approach: Luigi has worked on a formalisation of the authoritativeness in terms of annotations and annotated programs.
- We discussed how the ranking metrics could be integrated in the reasoning framework: we initially discussed using the ranks and constraints to block inferences, but this turned out to be infeasible in practice. Instead, ranks are now "calculated" for inferences using a min aggregation of the ranks of the triples satisfying the body of the rule. Along with the input triple ranks, these ranks can provide useful information for subsequent post-processing of inconsistencies.

- We eliminated some possible approaches which we had initially intended to apply, but which on further analysis we found to be practically or theoretically undesirable: for example, we ruled out using contra-positives.
- We performed some initial experiments over real Web data: although preliminary, we found some encouraging examples in the data for which our initial approach would work well.
- We agreed upon a rough schedule of work which would allow us to work in parallel on separate tasks: Pierro and Luigi will work on the underlying formalisations whilst in Galway, Axel and I pursue implementation and experimentation. Our initial aim is to submit to “JWS special issue on Provenance and Semantic Web”. The time-frame may prove prohibitively short for this venue, but we find the CFP particularly attractive for our work.

Although a brief STSM, we found the time particularly productive in terms of ironing out the details of the collaboration we had previously agreed upon, and for deriving a future plan of progression along those lines. Again, we hope to presently publish the results of our collaboration in a conference and/or journal venue.

References

[1] Aidan Hogan, Andreas Harth, Axel Polleres. "Scalable Authoritative OWL Reasoning for the Web". In International Journal on Semantic Web and Information Systems (IJSWIS), 5(2), April-June, 2009.

[2] Aidan Hogan, Andreas Harth, Jürgen Umbrich, Stefan Decker. "Towards a Scalable Search and Query Engine for the Web". Poster at 16th International World Wide Web Conference. System online at http://swse.deri.org

[3] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, Axel Polleres. "Weaving the Pedantic Web". In Proceedings of the Linked Data on the Web WWW2010 Workshop (LDOW 2010), Raleigh, North Carolina, USA, 27 April, 2010.