Towards a theoretical foundation for the harmonization of linked data

Enrico Daga*

Knowledge Media Institute (The Open University), United Kingdom
Semantic Technology Laboratory (ISTC, Consiglio Nazionale delle Ricerche), Italy

Abstract. In real world cases, building *reliable problem centric views* over Linked Data [1] is a challenging task. An ideal method should include a formal representation of the requirements of the needed dataset and a controlled process moving from the original sources to the outcome. We believe that a goal oriented approach, similar to the AI planning problem, could be successful in controlling the process of linked data fusion, as well as to formalize the relations between requirements, process and result.

Keywords: Linked Data, Data Harmonization, Planning

1 Introduction

We intend *Linked Data Harmonization* to be a controlled data preparation process, which transforms, aggregates, filters, fix and clean information from various linked data sources into a new *harmonized dataset* to fulfill the needs of a specific problem space, expressed as a formalized data schema. There is no single command that can achieve this. In contrast several actions must be performed in order to reach the goal (SPARQL, Linking, Programming, Rules, Reasoning, etc.). This task strongly resembles the AI planning problem. A planner takes a goal, a description of object types and properties as well as possible actions, a description of the initial state of the world, and returns as output a sequence of actions that will achieve the goal, when executed. The hypothesis we wish to verify is the following: *we can define a theory for the definition of plans for the integration of linked data whose accuracy is verifiable with respect to the needs of a particular task.*

Recently, a serious evaluation of the reliability of the linked data paradigm is emerging. This includes discussions about the capabilities of the tools for exploiting linked data [2] as well as on how this information could be effectively reused for reliable data analysis task [3].

We classify the methods for linked data integration in two categories:

(1) *goal/query oriented*: the user specifies a set of requirements in a declarative way (the query language) and data is kept where it actually is, the integration being factored at query processing level (for example [4]). This approach

* A special thank to Angelo Oddi (Planning and Scheduling Team PST, ISTC-CNR) which introduced me to the planning and helped me in the design of the experiment.

formalizes the requirements (the query) but the reliability of its output depends on the real-time availability of remote systems and on the limitations of the capabilities of the query language (with respect to entity linking, for example).

(2) *process/data oriented*: a user customizes a (set of) tools in order to define a process to build a dataset from the sources able to answer a set of (implicit) domain requirements (an example is LDIF [5]). This approach makes feasible to combine multiple commands for dealing with sub-tasks (like the linking of equal entities), but does not provide a way to formally express requirements and goals.

Other approaches include the formalization of prototypical tasks through reasoning patterns [6], approaching the semantic web as a unique ontology and not as linked data, and the field of ontology matching [7], which focuses on the ontological level (the OWL level) instead of the data structure (the RDF level). Existing approaches however do not deal with the problem of specifying data requirements and ensuring reliability with respect to these requirements. A goal/data oriented approach as the solution we envisage here seems not to be addressed by existing methods and tools.

2 Towards a theory for Linked Data Harmonization

2.1 Methodology

To formulate our theory we consider four tasks[1].

1. Represent a dataset and its portions. We base our model on the concepts of `Dataset`, graph `Slice` - a *pattern* for detecting a coherent subsets of triples according to some criteria and `Symbol`, representing any RDF resource. In VoID [8] the concepts of property and class partition have been introduced, while [9] used the concept of Path - all are kind of slices in our model.

2. Model properties and operators. Properties describe the features of a dataset, or of a specific slice. For example, a property may indicate the presence of a given slice in a dataset or describe relations between symbols. For example two predicates are reversible or another one may represent the amount of values for a predicate on any subject. *Operators* encode the actions that can be performed on a dataset, for example `COPY`, `FILTER`, `APPEND`. They have *parameters* and *effects*. Parameters bind the functionality to graph properties (*preconditions*), which constrain the operator to be applicable on a specific dataset state. Effects are consequences of the executed action described in term of dataset properties. Dataset model, properties and operators constitute the planning *domain*, which encodes type of objects and possible actions involving them.

3. Model requirements. This is a description of the initial state and of the goal. The *goal* is the expression of the task in terms of properties of the *goal dataset*, while the *initial state* includes the properties of the *source datasets* and the relations between the symbols used in both.

[1] Follows a synthetic description of each aspect. More details and relevant online resources are available at `http://www.enridaga.net/phd/iswc2012/`

4. Produce harmonization plans. We intend to simulate a data fusion process with a state of the art planner and evaluate how it may support our hypothesis (and to what extend). Then, if necessary, build our own tool for generating plans to be run by state of the art linked data frameworks, such as LDIF [5].

2.2 Evaluation

We intend to evaluate our theory and the resulting methodology in the following ways: (1) analyzing the class of harmonization situations it is able to support (a qualitative evaluation of our hypothesis); (2) doing a task based evaluation (how much effort is required with/without this approach in a given scenario? How much is the cost of interoperation of our data with data consuming tools?); (3) defining a scenario and manually executing the process using SPARQL, state of the art tools and by writing an ad-hoc program. The resulting dataset will be the *gold standard* to compare with the one produced by our tool. This should evaluate the overall approach from the user point of view.

It is a theoretical problem to understand how many real-world situations our theory may cover, so we intend to discuss also unsupported scenarios.

3 Lessons learnt from an initial experiment

We defined a pilot use case starting from the following exemplary task:

> *Report about the number of tenders from the EU in public infrastructures of a specific country along with the number of citizens living in the region.*

We identified 2 data sources: (1) LOTED [10] - which contains information about countries and tenders over the years (2) EUROSTAT (via ontologycentral.com), to retrieve statistics about population. A initial attempt we considered deterministic planning[2]. A subset of the theory have been encoded as PDDL[3] *domain*, and a requirements as *problem*. As test, the Fast Downward[4] planner has been used[5]. This experiment allowed us to do a first evaluation of the feasibility of the approach. We have been able to discover a valid plan. However we needed to make several compromises in the modeling phase. There is a trade off between computational efficiency (computability) and expressivity of the domain. To make the planner find a plan, we needed to have exactly the properties, operators and objects useful to solve this single problem - nothing less and nothing more. In addition, we discovered several limitations of a classic deterministic

[2] For an overview of deterministic planning and recent advances in the field, see [11].
[3] The Planning Domain Definition Language, which is the de facto standard for describe the features of a deterministic planner [11]
[4] http://www.fast-downward.org/
[5] PDDL files and problem solution are available at `http://www.enridaga.net/phd/iswc2012`.

planner that we started to analyze taking PDDL as reference specification: a data integration process needs to clone, create and destroy objects (datasets are appended, slices are copied); slices have several complex relations (any slice contains potentially many others, and it could be necessary to know if a needed slice can be obtained by specializing an available one); initial knowledge can be uncertain (the planner should be able to inspect available graphs on demand): all these features are not supported by PDDL. The following steps are to complete the analysis of the requirements a planner must satisfy in order to support our theory and to implement a software able to solve harmonization problems.

References

1. T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.
2. C.R. Rivero, A. Schultz, C. Bizer, and D. Ruiz. Benchmarking the performance of linked data translation systems. In *Linked Data on the Web Workshop at WWW 2012*, 2012.
3. S. Auer. Creating knowledge out of interlinked data: making the web a data washing machine. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, New York, NY, USA, 2011. ACM.
4. J. Álvarez, J. Labra, R. Calmeau, Á. Marín, and J. Marín. Query expansion methods and performance evaluation for reusing linking open data of the european public procurement notices. *Advances in Artificial Intelligence*, 2011.
5. A. Schultz, A. Matteini, R. Isele, C. Bizer, and C. Becker. Ldif-linked data integration framework. In *2nd International Workshop on Consuming Linked Data, Bonn, Germany*, 2011.
6. F. Van Harmelen, A. Ten Teije, and H. Wache. Knowledge engineering rediscovered: towards reasoning patterns for the semantic web. In *Proceedings of the fifth international conference on Knowledge capture*, pages 81–88. ACM, 2009.
7. P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics IV*, pages 146–171, 2005.
8. K. Alexander and M. Hausenblas. Describing linked datasets-on the design and usage of void, the vocabulary of interlinked datasets. In *In Linked Data on the Web Workshop (LDOW 09), International World Wide Web Conference*, 2009.
9. V. Presutti, L. Aroyo, A. Adamou, B. Schopman, A. Gangemi, and G. Schreiber. Extracting core knowledge from linked data. In *The 2nd Int. Workshop on Consuming Linked Data (COLD 2011) at ISWC 2011*, 2011.
10. F. Valle, M. dAquin, T. Di Noia, and E. Motta. Loted: Exploiting linked data in analyzing european procurement notices. In *Proceedings of the 1st EKAW Workshop on Knowledge Injection into and Extraction from Linked Data*, 2010.
11. A.E. Gerevini, P. Haslum, D. Long, A. Saetti, and Y. Dimopoulos. Deterministic planning in the fifth international planning competition: Pddl3 and experimental evaluation of the planners. *Artificial Intelligence*, 173(5-6):619–668, 2009.