Mapping, enriching and interlinking data from heterogeneous distributed sources

Anastasia Dimou
supervised by Rik Van de Walle, Erik Mannens, and Ruben Verborgh

Ghent University iMinds Multimedia Lab
Gaston Crommenlaan 8 bus 201, 9050 Ghent, Belgium
anastasia.dimou@ugent.be

Abstract. As Linked Open Data is gaining traction, publishers incorporate more their data to the cloud. Since the whole Web of Data cannot be semantically represented though, data consumers should also be able to map any content to RDF *on-demand* to answer complicated queries by integrating information from multiple heterogeneous sources distributed over the Web or not. In both cases, the quality and integrity of the generated RDF output affects the performance of traversing and querying the Linked Open Data. Thus, well-considered and automated approaches to semantically represent and interlink, already during mapping, the domain level information of distributed heterogeneous sources is required. In this paper, we outline a plan to tackle this problem: We propose a uniform way of defining how to map and interlink data from heterogeneous sources, alternative approaches to perform the mappings and methods to assess the quality and integrity of the resulting Linked Data sets.

1 Problem Statement

Efficiently extracting and integrating information from diverse, distributed and heterogeneous sources to enable rich knowledge generation that can more accurately answer complicated queries and lead to effective decision making, remains one of the most significant challenges. Nowadays, Semantic Web enabled technologies become more mature and the RDF data model is gaining traction as a prominent solution for knowledge representation. However, only a limited amount of data is available as Linked Data, because, despite the significant number of existing tools, acquiring its RDF representation remains complicated.

Deploying the five stars of the Linked Open Data schema[1] is still the de-facto way of incorporating data to the Linked Open Data (LOD) cloud. Approaching though the stars as a set of consecutive steps and applying them to separately individual sources, disregarding possible prior definitions and links to other entities, leads in failing to reach the uppermost goal of publishing interlinked data. Manual alignment to their prior appearances is often performed by redefining their semantic representations, while links to other entities are defined after the

[1] http://5stardata.info/

data is mapped and published. Identifying, interlinking or replicating, and keeping them aligned is complicated and the situation aggravates the more data is mapped and published. Existing solutions tend to generate multiple Unique Resource Identifiers (URIs) for the same entities while duplicates can be found even within a publisher's own datasets. Hence, demand emerges for a well-considered policy regarding mapping and interlinking of data in the context of a certain knowledge domain, either to incorporate the semantically enriched data to the LOD or to answer a query on-the-fly.

So far, there is neither uniform mapping formalisation to define how to map and interlink heterogeneous distributed sources into RDF in an integrated and interoperable fashion nor complete solution that supports the whole mapping and interlinking procedure together. Apart from few domain specific tools, none of the existing solution offer the option to automatically detect the described domain and propose corresponding mapping rules. Except for the field of plain text analysis where again the main focus is on semantically annotating the text rather than describing a domain and the relationships between its entities. Moreover, there are are no means to validate and check the consistency, quality and integrity of the generated output, apart from manual user-driven controls, and no means to automate these tests and incorporate them in the mapping procedure.

2 Relevancy

The problem is directly relevant to *data publishing* and *data consumption* with an emphasis on *semantically-enabled data integration*. In the *data publishing* end of spectrum, domain level information can be integrated from a combination of heterogeneous sources and published as Linked Data, using the RDF data model. In the *data consumption* end of spectrum, the relevancy is two-fold: (i) On the one hand, the quality and integrity of the resulting RDF representation is reflected at the dataset' s consumption. (ii) On the other hand, data extracts can be mapped and interlinked *on-demand* and *on-the-fly* from different heterogeneous sources, since not all data can be represented as Linked Data. On the whole, the problem is relevant to the alignment and synchronisation of data's semantic and non-semantic representations; modifications (inserts, updates and deletions) need to be synchronised over data's semantic and non-semantic representation.

The problem is emphasized in cases of knowledge acquisition, searching or query answering that information integration is required from a combination of distributed and heterogeneous (semantic and/or non-semantic) data sources. Especially when it is taken into consideration data that cannot be easily traversed else, for instance the deep Web or large volumes of published data files. Semantic Web technologies together with the RDF data model allows to deliberately concatenate the extract of data that is relevant.

There are several stakeholders that could take advantage of such information integration enhanced with semantic annotation. Such key stakeholders are those who publish and consume large volumes of data that might be distributed and appear in heterogeneous formats. For instance, governments that publish and

consume, at the same time, Open Data, scientists that combine data from different sources and re-publish processed information or (data) journalists that need extracts of data from several sources to acquire knowledge and draw conclusions.

3 Related Work

Several solutions exist to execute mappings from different file structures and serialisations to RDF. Different mapping languages beyond R2RML were defined [6] in the case of relational databases and several implementations already exist[2]. Similarly, mapping languages were defined to support conversion from data in CSV and spreadsheets to the RDF data model. For instance, the XLWrap's mapping language [10] that converts data in various spreadsheets to RDF, the declarative OWL-centric mapping language Mapping Master's M2 [11] that converts data from spreadsheets into the Web Ontology Language (OWL), Tarql[3] that follows a querying approach and Vertere[4] that follows a *triple-oriented* approach as R2RML does too. The main drawback in the case of most *row-oriented* mapping solutions is the assumption that each row describes an entity (*entity-per-row assumption*) and each column represents a property.

A larger variety of solutions exist to map data from XML to RDF, but to the best of our knowledge, no specific languages were defined for this, apart from the W3C standardized GRDDL[5] that essentially provides the links to the algorithms (typically represented in XSLT) that maps the data to RDF. Instead, tools mostly rely on existing XML solutions, such as XSLT (e.g., Krextor [9] and AstroGrid-D[6]), XPATH (e.g., Tripliser[7]), and XQUERY (e.g., XSPARQL [1]).

In general, most of the existing tools deploy mappings from a certain source format to RDF (*per-source approaches*) and only few tools provide mappings from *different* source formats to RDF. Datalift [12], The DataTank[8], Karma[9], OpenRefine[10], RDFizers[11] and Virtuoso Sponger[12] are the most well-known. But those tools actually either employ separate *source-centric* approaches for each of the formats they support, for instance Datalift, or rely on converting data from other formats to a *master* which in most cases is *table-structured*, for instance *Karma* or *Open Refine*. Furthermore, none of them provides an approach where the mapping definitions can be detached from the implementation.

Beyond pure execution of mappings to RDF, most of the existing tools do not provide any recommendations regarding how the data should be mapped,

[2] http://www.w3.org/2001/sw/rdb2rdf/wiki/Implementations
[3] https://github.com/cygri/tarql
[4] https://github.com/knudmoeller/Vertere-RDF
[5] http://www.w3.org/TR/grddl/
[6] http://www.gac-grid.de/project-products/Software/XML2RDF.html
[7] http://daverog.github.io/tripliser/
[8] http://thedatatank.com
[9] http://www.isi.edu/integration/karma/
[10] http://openrefine.org/
[11] http://simile.mit.edu/wiki/RDFizers
[12] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger

namely how to model the domain described. Only Karma offers mapping recommendation, however it relies on a training algorithm that improves when several domain-relevant data sources are mapped. Among the other tools, only Open Refine supports recommendations to a certain extend, but its recommendations have the form of disambiguating named entities appearing in the LOD.

As described, existing tools are solely focused on mapping data to the RDF model, rather than interlinking the entities of the source to existing entities appearing on the Web. Only Open Refine allows to reconcile and match entities to resources published as Linked Data and Datalift which incorporates interlinking functionality but only as a subsequent step executed after the mapping is completed. Overall, till nowadays, mapping and interlinking are considered two steps that are executed consecutively. A lot of work has been done in the field of text analysis, natural language processing (NLP) and named entity recognition (NER) to identify and disambiguate entities with resources appearing in the LOD cloud. However such techniques are mainly focused on semantically annotating the text rather than modelling the domain described. Moreover these techniques are not applied in the case of (semi-)structured mappings.

Last but not least, none of the existing tools offer a complete solution that allows to refine the executed mappings based on the users' feedback, the results of data cleansing tools, reasoning over the ontologies used or studying the integrity and connectedness of the resulting dataset considering it as a graph. A summary of existing approaches for assessing data quality that could be incorporated for refining the mappings according to the result of a mapping can be found at [13]. Among the pioneer tools for RDF data cleansing are the user-driven TripleCheckMate [8] and the test-driven RDFUnit [7]. Again only Karma is capable of refining its proposed mapping according to users' intervention.

4 Research Questions

The main question in my doctoral research is:

- *How can we access and represent domain level information from distributed heterogeneous sources in an integrated and uniform way?*

On the one hand, the accessing aspect needs to be investigated:

- *How can we enable querying distributed heterogeneous sources on the Web in a uniform way?*

On the other hand, the representation aspect needs to be investigated:

- *How can we identify if entities of a source have already been assigned a URI and enrich this unique representation with new properties and links?*
- *How can we interlink newly generated resources with existing ones already during mapping considering the available domain information we have?*

And the overall result raises the following questions:

- *How can we assure that if we map some sources the domain is accurately modelled?*
- *How well the entities of the dataset are linked with each other?*
- *How well the dataset is linked with the* LOD *cloud?*

5 Hypotheses

The main hypotheses related to my research are:

- *Integrated mapping and interlinking of data in heterogeneous sources generates fewer overlapping entities and models better the domain's semantics.*
- *Reusing Unique Resource Identifiers (*URIs*) leads to more robust and uniform datasets that have higher integrity and connectedness.*
- *Interlinking such datasets raises the integrity and connectedness of the whole* LOD *and improves the performance of its consumption.*
- *Not all media can be published as Linked Open Data, thus mapping extracts of multiple heterogeneous data to* RDF *might occur on demand.*

6 Approach

At this PhD, we propose a generic mapping methodology, that maps the data independently of the source structure (*source-agnostic*), puts the focus on mappings and their optimal reuse and considers interlinking already during mappings. Therefore, the initial learning costs remain limited, the potential for the custom-defined mapping's reuse augments and a richer and more meaningful interlinking is achieved. This is a prominent advancement compared to the approaches followed so far. As a result, the per-source mapping model followed so far gets surpassed, leading to contingent data integration and interlinking. Beyond the language that facilitates the mapping rules' definition and is the core of our solution, we propose a complete approach that aims to facilitate and improve the mappings definition and execution.

In our proposed approach we aim to maximize the reuse of existing unique identifiers (URIs) and rely on the links between them and the newly generated entities to achieve the interlinking of the new dataset with the LOD. The disambiguated entities are assigned the corresponding URIs and their representation is enriched with properties and relationships of the newly incorporated dataset. In contrast to the approaches followed so far, custom-generated URIs are only assigned to the entities that were not identified in the LOD cloud (not disambiguated). Based on the relationships between the newly generated entities and the disambiguated ones, the interlinking of the newly generated resources with the LOD is achieved. In order to identify such entities, we propose applying NER techniques to the sources and use them against datasets of the LOD.

Besides increasing the integrity of the dataset and reinforcing its interlinking with the LOD cloud, the whole domain needs to be modelled. Recommendations based on vocabularies used for the description and for the relationships of the disambiguated entities or other entities that are identified to model the same domain and those appearing in a vocabularies' repository, such as LOV[13], can be taken into consideration. The domain can be further refined after the execution of the mappings and the assessment of the output dataset using tools for evaluating

[13] http://lov.okfn.org

the data quality or taking into considerations the users' feedback. In these cases, the mapping rules can be adjusted to incorporate the emerging rules.

7 Preliminary results

We already defined a generic language adequate for defining rules to map heterogeneous sources into RDF in a uniform and integrated way [3]. This language is the RDF Mapping Language (RML) [14], defined as a superset of the W3C standardized mapping language R2RML. RML broadens R2RML's scope and extends its applicability to any data structure and format. RML came up as a result of our need to map heterogeneous data to RDF. Initially, R2RML was extended to map data from hierarchically structured sources e.g., XML or JSON, to RDF. Details about how we extended the row-oriented R2RML to deal with hierarchy, and other structures in general, are described in detail at our previous work [5].

Even though the language's extensibility is self-evident as RML relies on an extension over R2RML, its scalability was also proven by further extending it to map data published as HTML pages to the RDF data model. Results of the mappings from HTML to RDF using RML were presented at the Semantic Web publishing challenge of the 11th Extended Semantic Web Conference (ESWC14) [2]. At the moment, in total, RML and the prototype processor support, but are not limited, mappings from data in CSV, XML, JSON and HTML to the RDF data model.

A prototype processor[15] was designed and implemented as a *proof-of-concept* to accompany the RML mapping language. As RML extends R2RML, the processor is implemented using an existing open-source R2RML processor[16]. The RML processor was designed to have a modular architecture where the extraction and mapping modules are independently executed and the extraction module can be instantiated depending on the possible inputs. Short discussion regarding alternative approaches for processors supporting RML were discussed at [5].

Finally, some preliminary work on mapping rules' refinements by incorporating data consumers' feedback was presented at [4]. We showed how provenance generated during mapping can be used later on to identify the mapping rules that should be adjusted to incorporate data consumers' feedback.

8 Evaluation plan

There are different aspects of the proposed solution which need to be assessed and we are aiming to evaluate: the RML mapping language itself, the semantic annotations and the entities interlinking, the quality and integrity of the resulting dataset and the performance of the mapping execution.

– the language's potential in regard to (i) the range of input sources supported and their possible combinations for providing integrated mappings, namely

[14] http://rml.io
[15] https://github.com/mmlab/RMLProcessor
[16] https://github.com/antidot/db2triples

the language's *scalability* and *extensibility*; (ii) the language's *expressivity*, namely the coverage of possible alternative mapping rules, mainly in comparison to other languages (or approaches) and (iii) last, how *reusable* and *interoperable* the mapping descriptions are.
- the *validity*, *consistency* and *relevance* (especially when the domain is modelled according to automated recommendations) of the vocabularies used by the mapping rules to describe the domain knowledge.
- the *quality* of the output. To achieve this, both automated solutions assessing data quality and domain experts will be used to evaluate the resulting dataset in regard to the identified or generated entities, the provided semantic annotations, the interlinking and the overall modelling of the domain.
- the *accuracy* and the *precision* and *recall* of the retrieved, identified and enriched entities in conjunction with the *confidence* for the interlinked entities.
- the *integrity* of the resulting dataset and the overall *analysis* of the output's datasets in respect to its graph-based representation, for instance in and out degree, its connectivity, its density, bridges, paths etc.
- the *impact* of the resulting dataset's structure and interlinking in respect to its subsequent *consumption*. To be more precise, how *traversing* and *querying* the dataset is affected by the choices taken while modelling the knowledge domain. In the case of querying, we aim to examine both the *complexity* of the queries definition and the *time* and *overload* to execute them.
- finally, while the *performance* is important to verify that the mappings can be executed in reasonable time, the performance of an RML processor is not the main focus of this work. However, the two fundamental ways of executing the mappings (mapping-driven or data-driven) will be evaluated and compared to identify best use-cases. The execution planning of the mapping rules though is more interesting and will be deeper investigated and evaluated.

9 Reflections

The main difference of our approach compared to existing works on mapping data is that we (i) introduce the idea of a uniform way of dealing with the mapping of heterogeneous sources and (ii) introduce the aspect of interlinking while we perform the mapping of data to the RDF data model. We approach the mapping from a domain modelling perspective where the data is either incorporated to a partially described domain or is mapped combined, forming their own domain. This way, we achieve generating datasets with higher integrity that are already interlinked among each other and with the LOD and thus we reduce the effort for subsequent interlinking of resources and offer better conditions for their subsequent consumption.

Acknowledgement

The research described in this paper is funded by Ghent University, the Flemish Department of Economy, Science and Innovation (EWI), the Institute for the

Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

References

1. S. Bischof, S. Decker, T. Krennwallner, N. Lopes, and A. Polleres. Mapping between RDF and XML with XSPARQL. *Journal on Data Semantics*, 1(3):147–185, 2012.
2. A. Dimou, M. Vander Sande, P. Colpaert, L. De Vocht, R. Verborgh, E. Mannens, and R. Van de Walle. Extraction and semantic annotation of workshop proceedings in HTML using RML. In *Semantic Publishing Challenge of the 11th Extended Semantic Web Conference*, May 2014.
3. A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. RML: A generic language for integrated RDF mappings of heterogeneous data. In *Proceedings of the 7th Workshop on Linked Data on the Web*, Apr. 2014.
4. A. Dimou, M. Vander Sande, T. De Nies, R. Verborgh, E. Mannens, and R. Van de Walle. RDF mapping rules refinements according to data consumers feedback. In *2nd International World Wide Web Conference, Poster Track Proceedings*, 2014.
5. A. Dimou, M. Vander Sande, J. Slepicka, P. Szekely, E. Mannens, C. Knoblock, and R. Van de Walle. Mapping hierarchical sources into RDF using the RML mapping language. In *Proceedings of the 8th IEEE International Conference on Semantic Computing*, 2014.
6. M. Hert, G. Reif, and H. C. Gall. A comparison of RDB-to-RDF mapping languages. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 25–32. ACM, 2011.
7. D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 747–758. International World Wide Web Conferences Steering Committee, 2014.
8. D. Kontokostas, A. Zaveri, S. Auer, and J. Lehmann. Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In *Knowledge Engineering and the Semantic Web*, volume 394 of *Communications in Computer and Information Science*, pages 265–272. Springer Berlin Heidelberg, 2013.
9. C. Lange. Krextor - an extensible framework for contributing content math to the Web of Data. In *Proceedings of the 18th Calculemus and 10th international conference on Intelligent computer mathematics*, MKM'11. Springer-Verlag, 2011.
10. A. Langegger and W. Wöß. XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 359–374. Springer-Verlag, 2009.
11. M. J. O'Connor, C. Halaschek-Wiener, and M. A. Musen. Mapping Master: a flexible approach for mapping spreadsheets to OWL. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part II*, ISWC'10, pages 194–208. Springer-Verlag, 2010.
12. F. Scharffe, G. Atemezing, R. Troncy, F. Gandon, S. Villata, B. Bucher, F. Hamdi, L. Bihanic, G. Képéklian, F. Cotton, J. Euzenat, Z. Fan, P.-Y. Vandenbussche, and B. Vatant. Enabling Linked Data publication with the Datalift platform. In *Proc. AAAI workshop on semantic cities*, 2012.
13. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked open data: A survey. Submitted to the Semantic Web Journal., 2013.