Toward a View-oriented Approach for Aligning RDF-based Biomedical Repositories

A. Anguita¹; M. García-Remesal¹; D. de la Iglesia¹; N. Graf²; V. Maojo¹

¹Biomedical Informatics Group, Universidad Politécnica de Madrid, Madrid, Spain; ²Department of Paediatric Oncology and Haematology, Saarland University Hospital, Homburg, Germany

Keywords
Ontology alignment, database integration, RDF

Summary
Introduction: This article is part of the Focus Theme of Methods of Information in Medicine on "Managing Interoperability and Complexity in Health Systems".

Background: The need for complementary access to multiple RDF databases has fostered new lines of research, but also entailed new challenges due to data representation disparities. While several approaches for RDF-based database integration have been proposed, those focused on schema alignment have become the most widely adopted. All state-of-the-art solutions for aligning RDF-based sources resort to a simple technique inherited from legacy relational database integration methods. This technique – known as element-to-element (e2e) mappings – is based on establishing 1:1 mappings between single primitive elements – e.g. concepts, attributes, relationships, etc. – belonging to the source and target schemas. However, due to the intrinsic nature of RDF – a representation language based on defining tuples < subject, predicate, object > –, one may find RDF elements whose semantics vary dramatically when combined into a view involving other RDF elements – i.e. they depend on their context. The latter cannot be adequately represented in the target schema by resorting to the traditional e2e approach. These approaches fail to properly address this issue without explicitly modifying the target ontology, thus lacking the required expressiveness for properly reflecting the intended semantics in the alignment information.

Objectives: To enhance existing RDF schema alignment techniques by providing a mechanism to properly represent elements with context-dependent semantics, thus enabling users to perform more expressive alignments, including scenarios that cannot be adequately addressed by the existing approaches.

Methods: Instead of establishing 1:1 correspondences between single primitive elements of the schemas, we propose adopting a view-based approach. The latter is targeted at establishing mapping relationships between RDF subgraphs – that can be regarded as the equivalent of views in traditional databases –, rather than between single schema elements. This approach enables users to represent scenarios defined by context-dependent RDF elements that cannot be properly represented when adopting the currently existing approaches.

Results: We developed a software tool implementing our view-based strategy. Our tool is currently being used in the context of the European Commission funded p-medicine project, targeted at creating a technological framework to integrate clinical and genomic data to facilitate the development of personalized drugs and therapies for cancer, based on the genetic profile of the patient. We used our tool to integrate different RDF-based databases – including different repositories of clinical trials and DICOM images – using the Health Data Ontology Trunk (HDOT) ontology as the target schema.

Conclusions: The importance of database integration methods and tools in the context of biomedical research has been widely recognized. Modern research in this area – e.g. identification of disease biomarkers, or design of personalized therapies – heavily relies on the availability of a technical framework to enable researchers to uniformly access disparate repositories. We present a method and a tool that implement a novel alignment method specifically designed to support and enhance the integration of RDF-based data sources at schema (metadata) level. This approach provides an increased level of expressiveness compared to other existing solutions, and allows solving heterogeneity scenarios that cannot be properly represented using other state-of-the-art techniques.
1. Introduction

The availability of methods and tools to uniformly access multiple data sources, conversely to the classic single-database research approach, has led researchers to significant achievements, such as, for instance, identifying target proteins for fighting specific diseases [1], generating prediction models for infectious diseases [2], and other findings in the field of biomarker research [3, 4]. On the other hand, this multidatabase research paradigm entails new challenges for biomedical informaticians, due to the difficulties that arise due to the disparity in the representation of information across different data sources [5].

These difficulties have been partly solved by the development of RDF (http://www.w3.org/RDF), which has been widely used over the last few years for modeling biomedical data. Furthermore, recent research reports [6] that RDF-based access to biomedical databases facilitates solving both technical and syntactic heterogeneities, thus allowing researchers to concentrate on solving semantic heterogeneities, a much harder – and still open – research problem.

Most existing database integration tools address the semantic integration task at schema level, by aligning the (semantically) heterogeneous schemas/metadata describing the sources to be integrated. The semantic integration can be achieved either by directly linking similar entities belonging to the different schemas to be integrated, or by aligning each individual schema to a common conceptualization, usually formalized as a domain ontology or a terminology. In the latter case, the semantic integration is achieved by linking semantically similar elements belonging to the schemas of the sources to be integrated to the same object (or objects) in the ontology. In this paper, we will focus on the latter case.

An advantage of systems that resort to a domain ontology to perform the semantic reconciliation is that the ontology itself can be used for querying the integrated set of repositories. On the other hand, the downside of this approach is that a suitable domain ontology is required. Therefore, much effort has been put over the last years on the development of RDF-based ontologies to describe specific domains of knowledge. These developed ontologies can then be used for solving the existing semantic heterogeneities between disparate sources [7].

The use of common RDF-based ontologies or vocabularies in semantic homogenization processes involves generating formal descriptions of the semantic relations holding among objects from the source schemas and those belonging to the target schemas [8]. These formalizations are commonly referred to as “mappings” in the database integration jargon. Mappings are commonly defined in a element-to-element (e2e) basis for most traditional semantic integration methods – including those focused on relational databases. In e2e-based approaches, mapping relationships are specified as links between the primitive representational elements belonging to the physical schemas of the databases to be integrated and those primitive representation elements belonging to the schema of the “virtual” or integrated database.

For instance, when applying the e2e approach to integrate relational databases, primitive representational elements are tables or fields – depending on the required level of granularity. In this scenario, e2e-based alignments are formalized as sets of table-to-table or field-to-field mappings. Existing systems reported in the literature for bringing together relational databases use different atomic representational elements. For instance, the OntoFusion system [9] developed by the authors resorts to the field as the atomic representational element, and thus, schema alignments are represented as sets of field-to-field mappings. Other systems, such as TAMBIS [10] or OBSERVER [11] work at a higher level of granularity, and thus resort to table-to-table e2e-based mappings.

Regarding the integration of RDF-based sources, the e2e-based approach has also been used in the vast majority of state-of-the-art methods and tools. We believe, however, that the e2e approach is not completely suitable for undertaking this task, due to the unique formalism used in RDF paradigm for representing and storing information. RDF sources are built upon the concept of “triple” rather than on tables or fields. Triples are ternary tuples <subject, predicate, object> which represent meaningful statements – e.g. “Patient is Treated By Clinician”. RDF-based models are built by combining multiple triples into graphs composed by RDF classes, properties and datatypes. In this context, there exist RDF elements – e.g. RDF classes, properties or data types – that depend on other RDF elements included in the same RDF subgraph to acquire the intended semantics.

This context-enabled semantic acquisition cannot be properly reflected by the e2e methodology, since the latter is limited to performing field-to-field or class-to-class mappings. This handicap neglects the possibility of embedding the required context-based mapping information into the alignment data without modifying the target schema – i.e. adding new classes, properties and data types. However, modifying a widely recognized ontology developed and curated by domain experts is not a recommended practice. Instead, when bringing together RDF-based data sources, the original alignment methodology must be extended to support not only e2e mapping relationships, but also alignments among views belonging to both the source and the target schema. Therefore, our proposal resorts to a view-oriented approach for RDF-based repository alignment: an alignment between two arbitrary RDF schemas should define pairs of semantically equivalent views from the two schemas instead of e2e mapping relationships between single RDF elements. In other words, the chosen primitive representational element are no longer classes, properties or data types – and not even triples – but views.

As stated above, most existing tools for RDF alignment are limited to e2e mappings between single RDF elements – i.e. they perform class-to-class and relation-to-relation mappings –, and thus suffer from the issues reported above. To better illustrate such problems, we will refer to a real-world RDF-based repository integration scenario that we faced during our participation in the already completed European Commission-funded Advanced Clinico-Genomic Trials (ACGT) research project [12]. In this scenario, we aimed at aligning an
RDF source schema containing data regarding the treatments undergone by cancer patients to the ACGT Master Ontology (ACGT MO) [13]. The latter is an RDF-based shared conceptualization developed in the context of ACGT, developed in OWL and targeted at representing the domain of post-genomic clinical trials on Cancer — can be accessed at www.ifomis.org/act/1.0. In the source repository, two different RDF classes – BiopsyBefore and BiopsyAfter – are used to represent the results of biopsies carried out before and after chemotherapy treatments, respectively. The Patient class is linked to both the classes BiopsyBefore and BiopsyAfter via the has property. On the other hand, the ACGT MO resorts to a combination of RDF properties to describe the two BiopsyAfter and BiopsyBefore concepts. To relate a patient to her/his biopsy results, the Patient class is related to the Biopsy class through the undergoes property. To distinguish between biopsies occurred either before or after chemotherapy, the ACGT MO introduces another relationship that links Biopsy to Chemotherapy. The latter relationship is one of precedes or preceded by to indicate that the biopsy intervention is carried out either before or after the chemotherapy treatment, respectively. Therefore, to perform a complete alignment, we would have to create two different view-based sub-alignments. First, we would map the source view \{Patient, has, BiopsyBefore\} to the following target view: \{Patient, undergoes, Biopsy, Biopsy, precedes, Chemotherapy\}. Next, we would align the source view \{Patient, has, Biopsy\} to the following target view: \{Patient, undergoes, Biopsy\}. As can be seen, the source views have been properly represented in terms of an RDF sub-graph belonging to the target schema. If we attempted to align both schemas using a e2e based approach, we would soon realize that the only possible alignment would involve mapping the source Patient class to the target Patient class and both the BiopsyBefore and BiopsyAfter to the Biopsy class, thus failing to incorporate the semantics regarding to whether the biopsy was performed before or after chemotherapy. A possible solution to this issue using e2e mappings would involve creating the BiopsyAfter and BiopsyBefore classes as subclasses of the Biopsy class in the ACGT MO. However, modifying the ACGT MO’s structure and contents may invalidate previous alignments carried out against other different sources. As can be seen, our approach is flexible enough and does not require modifying the target ontology to perform complex alignments.

This paper is organized as follows. In section 3 we explain how to perform view-oriented alignments between the metadata describing the RDF-based sources to be integrated and a target RDF/OWL ontology – i.e. the common conceptualization that models the domain described by the “virtual” integrated database. In section 4 we present the software tool we have created to implement the view-oriented alignment strategy and we show how to implement a view-based alignment for the example introduced above. Section 5 presents a comparison between our proposal and the existing state-of-the-art solutions. Finally, section 6 draws the conclusions and points to future research lines.

2. A methodology for Defining Alignments between RDF-based Models

We have created a methodology for defining alignments between RDF-based models – we purposely use the term model to refer to RDF database schema/meta-data, RDF-based vocabularies or ontologies developed in RDF or any of its extensions, since our approach is applicable to any of these entities – in the context of the data integration layer of the ongoing EC-funded “p-medicine: from data sharing and integration via VPH models to personalized medicine” (p-medicine from now onwards) project [14]. This method is based on our view-oriented alignment approach, which defines mappings as pairs of semantically equivalent RDF views, each view belonging to a different RDF-based data model. Thus, our method receives as inputs a source schema – the metadata of one of the sources to be integrated – and a target schema – a domain ontology encoded in RDF or OWL –, and produces a set of pairs of RDF views encoding the semantic mappings.

Compared to classical e2e-based alignments used in existing solutions, our approach involves an additional level of complexity for establishing the semantic alignments. The process for defining each mapping is composed of several steps, as described below:

1. Build a view over the source schema representing a specific scenario or situation to be modeled. Domain experts can help identifying such relevant scenarios.
2. Define a view over the target schema that accurately reproduces – i.e. is semantically equivalent to – the same scenario as above.
3. Establishing e2e mappings between the individual RDF elements of the views built in the previous steps.
4. Repeat from step 1 for as many scenarios or situations as needed.

The result is a set of pairs of RDF views, which we call "entries". Each entry is composed of a couple of views – one from the source schema and another belonging to the target schema – and a set of e2e mappings among elements belonging to both views – such as those used in classical e2e-based alignment approaches.

2.1 View Definition and Organization

A view, as described in database theory, represents a subset of the schema that models the domain covered by a database. In RDF, representational primitives are tuples of two class instances linked by a relationship. RDF views are therefore defined as sets of RDF tuples that generate a sub-graph subsumed in the complete RDF model. Nevertheless, not all potential sub-graphs of the complete RDF model are considered. We will constrain views to be represented as connected graphs, since non-connected graphs do not provide any semantically consistent information.

The alignment information is represented in a XML-based mapping format described elsewhere [15]. To present mapping information to end users in a more natural way, we represented each view as a
set of paths. Paths can be defined as a sequence of alternate RDF classes and relations, beginning with a class and ending with either a class or an RDF datatype. Paths can be joined to each other at coincidental classes to form the required connected graph. Therefore, a view – both for the source and the target schemas – is composed of a set of connected RDF paths. For example, the view “Female patients that participate in a clinical trial” is decomposed into the pair of paths “Female patients” and “Patients that participate in a clinical trial”.

2.2 Translation Process

Our approach to view-oriented alignments between RDF models leads to a more straightforward translation process compared to the classical e2e based alignments. To translate the data from an RDF data source into concepts belonging to a target vocabulary – e.g. in case of adopting a centralized data integration approach consisting of several databases whose data contents are fed into a central Data Warehouse – the physical views of each entry in the alignment file describe how to extract the information from the respective database. Each RDF view is then translated into a SPARQL (http://www.w3.org/TR/rdf-sparql-query/) query, and the results are used to populate the target schema as defined in the corresponding conceptual view. The e2e mappings stored in each entry specify which elements from the physical database are extracted, and to which elements in the conceptual schema they are transferred.

3. A Web-based Tool for Building View-oriented Alignments

We developed a web-based tool with a graphical user interface (GUI) for enabling users to build view-oriented alignments between two arbitrary RDF schemas. Our tool provides several canvases aimed at both representing the aligned schemas and at building the RDF views for each of the two schemas. Figure 1 shows a screenshot of our tool’s main window.

Our tool was validated in the context of real scenarios provided by physicians in the p-medicine project. In these tests, we used the Health Data Ontology Trunk (HDOT) ontology [16] covering the domain of clinical trials on cancer as the target schema. We aligned three different RDF-based sources to HDOT using our tool. Two of these sources contained data from clinical trials on cancer – based on the BFM-ALL study [17] on acute lymphoblastic leukaemia cancer and retrospective data from breast cancer clinical trials, respectively. The third database contained DICOM images. Each alignment involved the construction of a set of entries, each containing one RDF view from the physical database and one RDF view from HDOT. These views included several “context-dependent” RDF elements, such as those described in the example from section 1. For this example, the conceptual view built for the biopsy performed before the chemotherapy treatment included two RDF paths: one with the structure Patient → undergoes → Biopsy, and other with Biopsy → precedes → Chemotherapy. This view was graphically constructed with our tool by building a sub-graph of the HDOT ontology, as shown on the lower-right canvas in Figure 1.

The resulting alignment files were used as input for the subsequent data translation process, which was capable of successfully translating the data stored in the physical databases into the common format provided by HDOT.
4. Discussion

There exist a plethora of tools for defining alignments between RDF models that can be found in the literature. Some of the most remarkable are COMA++ [18], Sambo [19], Snoggle [20], Optima [21], Ontostudio [22], AgreementMaker [23] and BioMixer [24]. These tools provide different features and capabilities, and resort to different data structures for representing the aligned RDF schemas – e.g. tree-like structures, graph-based representations, etc. However, all of them rely on e2e mappings between individual elements (classes, attributes or relations) belonging to both source and target schemas – a feature inherited from initial efforts for integrating relational databases, as stated previously. We believe that the e2e approach is suitable for naive scenarios that just involve simple situations – e.g. mapping of class attributes that do not involve additional relations to other classes –, but fail to properly support the alignment of “context-dependent” elements that we showed in the previous section without performing modifications to the target schema, that, on the other hand, should be avoided when possible.

Our view-based mapping approach can be regarded as a practical variation of the Global-Local-As-View (GLAV) approach [25] applied to alignments between RDF-based models. The GLAV language for mappings aims to solve the shortcomings of the Global-As-View (GAV) [26] and Local-As-View (LAV) [27] integration approaches and provides a solution to issues that arise due to the adoption of different levels of granularity and different database normalizations – the latter is essentially the problem that we aim at solving with our approach. There exist other works based on GLAV. For instance, Calvanese et al. introduced an algorithm for processing queries under GLAV mappings with relational databases [28]. Knoblock et al. developed KARMA, a standalone tool for generating GLAV mappings between relational sources and ontologies [29, 30]. Using KARMA, users can effectively map tabular data to an ontology view. We reproduced the experiment described in the previous sections using a tabularized version of our RDF source data, showing that KARMA is also capable of solving the same problem. This was possible because we managed to manually translate the RDF source to the tabular format required by KARMA. However, in other scenarios this manual translation of RDF data into tabular data – although feasible – might be a hard and time-consuming task. In such circumstances, our approach can natively handle the alignment of RDF-based sources. To sum up, KARMA successfully addresses the issues derived from using e2e mappings, but it is constrained to bridging relational sources or tabular data. Therefore, KARMA can be regarded as a complimentary tool to our proposal, since, despite that both problems share some similarities, they also present specific issues inherent to the relational and RDF models. Regarding the limitations of our proposal, our developed tool cannot yet entirely handle different levels of granularity. This is so, since our mapping format does not yet allow populating one target field with the information of more than one source field, and vice-versa. This issue is currently under research. The experiment described in the previous section presents a sample scenario that can be properly addressed following our approach but that cannot be handled by resorting to the classical e2e-based alignment approach without modifying the target schema. In this sample scenario it is not possible to simultaneously map the BiopsyBefore and BiopsyAfter source classes to the Biopsy target class in ACGT MO without losing the semantics regarding whether the biopsy was done before or after chemotherapy. This issue could be addressed by existing methods such as those mentioned above by extending ACGT MO with two different subclasses – e.g. BiopsyBefore and BiopsyAfter – of the Biopsy target class. However, as stated previously, independently modifying the target schema is not a recommended practice since this may break the homogeneity between this mapping and other previously defined alignments. In addition, the original ACGT MO representation of the BiopsyBefore and BiopsyAfter classes – provided by experts via the precedes and preceded by relationships – becomes now redundant and useless. Conversely, by resorting to our view-based strategy, it is possible defining two different paths (views) on the Biopsy and Chemotherapy classes to represent the times when the biopsies were carried out via the expert-provided ACGT MO precedes and preceded by relationships.

The proposed approach was initially developed to solve the semantic inconsistencies between the ACGT MO and several clinical databases during the ACGT project. In the follow-up project p-medicine, we are currently dealing with the semantic inconsistencies between a new cancer domain ontology (HDOT) and a set of clinical and genomic databases. In this context we found that the e2e mapping approach would not be expressive enough to formally describe the mappings between these databases and HDOT, due to differences in domain modeling or missing context information. In essence, this approach can be extended to work using any existing RDF-based ontology, as its use only depends on the modeling differences between the ontology and the sources, not to the ontology itself.

We believe that by following our proposed mapping strategy it is possible to properly align any source RDF path to a suitable target RDF view provided that this exists in the target ontology. This allows preserving the homogeneity between the current alignment and other previously defined mappings, since all alignments are made by resorting to expert-provided views already present in the target schema. This avoids introducing additional ad hoc RDF elements to the target schema.

5. Conclusions and Future Work

Schema alignment is a hot topic for most biomedical informaticians and computer scientists. Properly aligning different data sources is required to enable seamless access to heterogeneous repositories. As mentioned previously, there exist a number of tools for defining alignment information between different databases – in general, ontologies. However, these rely on the classical e2e mappings between RDF elements – to the best of our knowledge, no existing tool yet provides a different alignment solution. This approach, inherited
from the alignment of relational databases, is valid for simple scenarios, but cannot properly deal with the complexity of context-dependent elements inherent to the RDF paradigm. Our approach, based on a view-oriented solution for defining mappings among RDF-based data schemas, can be seen as a generalization of the e2e alignment approach, and provides the capability of modeling any type of alignment between RDF-based data sources.

The presented approach has been successfully implemented in a web-based tool for the alignment of RDF database schemas, tested with real data scenarios and used to translate data coming from different databases into terms of a common vocabulary.

Future work will be aimed at improving the usability of the tool by adding (semi)automated alignment procedures. Current automated procedures for discovering semantic relations between RDF-based sources are focused on finding e2e relations based on different heuristics such as lexical analysis or morphological relations—e.g. lexical/structural/semantic distances between pairs of terms. In our view-based approach, this is no longer valid. Instead, it will be necessary to generalize existing solutions to accommodate to the view-based mappings. The difficulty lies in the complexity growth. While in e2e mappings the number of possible mappings was limited to n · m, where n and m are the number of elements in each schema, in the view-based approach we must consider combinations of elements to form views—assuming only views formed by different elements, there are \((2^n - 1) \cdot (2^m - 1)\) combinations. A brute force solution is no longer feasible. Hence, it will be required to analyze more intelligent approaches to explore the whole space of potential mappings.

Acknowledgments

This project has received funding from the European Union’s Seventh Programme for research, technological development and demonstration under grant agreement No 270089.

References