DataOps: Seamless End-to-End Anything-to-RDF Data Integration

Christoph Pinkel(✉), Andreas Schwarte, Johannes Trame, Andriy Nikolov, Ana Sasa Bastinos, and Tobias Zeuch

Fluid Operations AG, Walldorf, Germany
{christoph.pinkel,andreas.schwarte,johannes.trame,
andriy.nikolov,ana.sasa,tobias.zeuch}@fluidops.com

Abstract. While individual components for semantic data integration are commonly available, end-to-end solutions are rare.

We demonstrate DataOps, a seamless Anything-to-RDF semantic data integration toolkit. DataOps supports the integration of both semantic and non-semantic data from an extensible host of different formats. Setting up data sources end-to-end works in three steps: (1) accessing the data from arbitrary locations in different formats, (2) specifying mappings depending on the data format (e.g., R2RML for relational data), and (3) consolidating new data with existing data instances (e.g., by establishing owl:sameAs links). All steps are supported through a fully integrated Web interface with configuration forms and different mapping editors. Visitors of the demo will be able to perform all three steps of the integration process.

1 Introduction

In recent years semantic data integration has evolved to an important application area in the industry: software eco systems in companies become more and more complex, produce large amounts of heterogenous information, and make it harder and harder to get a holistic view on the company's knowledge.

Traditionally, in such situations dedicated ETL-style systems are used for the analysis. Functionality is provided by large scale data warehouse systems or, more recently, by big data frameworks such as Hadoop YARN [1] that work as a *data operating systems* running a mix of data warehousing and other applications. Those systems share an important property for enterprise data analysis: available as ready solutions, they include everything - from assisted setup, a broad selection of access methods, over graphical configuration interfaces to a comprehensive documentation and support.

However, they come along with a downside: in classical data warehouses a dedicated, global warehousing schema must be designed, mappings must be constructed, and the resulting schema must be documented and communicated to users. For Hadoop-like systems, this is not necessarily the case, as a mix of relational and non-relational workloads are possible with different applications in the system. This comes at the price of either a very small set of supported

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9341, pp. 123–127, 2015.
DOI: 10.1007/978-3-319-25639-9_24

queries and little flexibility, or involves even more initial effort for programming all the tasks and queries to be supported. Worse, with a number of data sources that quickly change in structure, maintenance of the resulting schema, mapping and queries can quickly become a nightmare in either case.

Often enough, the effort for setup and maintenance becomes unacceptable, especially if some data sources are complex in structure. This contributes to the current situation where enterprises are assumed to analyze less than one sixth of their potentially relevant data.[1] Semantic data integration, with its flexible graph model and vocabularies, is one possible and natural way to address this predicament.

To leverage the potential of semantic data integration in such cases, an integration environment needs to be fully functional and sufficiently usable, much the same as enterprise data warehousing solutions. While individual components for semantic data integration are commonly available, end-to-end solutions that fulfill all requirements are rare. Existing frameworks such as the Linked Data Integration Framework [3] focus on web-scale data rather than enterprise data. Similarly, many systems composed of loosely coupled special-purpose components (where setup or maintenance involve intense human effort) fail the one key requirement that motivates their use in the first place: to significantly reduce overall effort in configuration and maintenance.

With DataOps, we target the challenge of delivering an end-to-end solution for semantic ETL data integration that supports seamless setup and configuration as well as convenient maintenance procedures. DataOps delivers all primarily relevant components as a toolkit out of the box, fills any gaps between them and offers an integrated user interface, built on industry-proven platform technology.

We describe details about the DataOps demonstration in Sect. 2 before we discuss selected related systems in Sect. 3. Finally, we conclude with the contributions of our demo (Sect. 4).

2 DataOps Demo

We demonstrate DataOps, a seamless Anything-to-RDF semantic data integration toolkit. DataOps supports the integration of both semantic and non-semantic data from a host of different formats, including relational databases, CSV, Excel, XML, JSON, existing RDF graphs and others. Additional source formats can be integrated through an extension mechanism. We demonstrate this with a specialized data source that allows to directly access results from the statistical software *R*. In addition, each data source can be accessed from different locations within an organization, e.g., as local files, from network shares (which may optionally require authentication), from the Web through HTTP, or even through custom protocols.

[1] According to a recent study of business analysts [2], surveying several hundreds of enterprises.

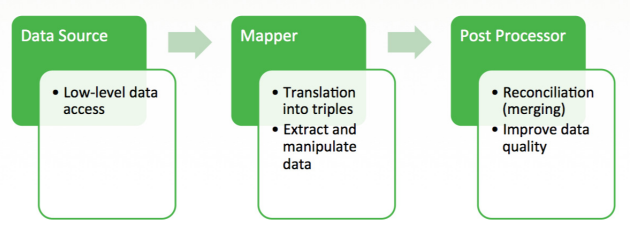


Fig. 1. DataOps process

As an integrated toolkit, DataOps supports all setup, configuration and maintenance steps through a fully integrated Web interface with configuration forms and different mapping editors. Setting up data sources end-to-end is implemented in a three-step process (Fig. 1):

1. Accessing the different data sources from arbitrary locations through different mechanisms (Fig. 2a).
2. Specifying mappings depending on the data source format (Fig. 2b).
3. Consolidating new data with existing data instances, e.g., by establishing owl:sameAs links (Fig. 2c).

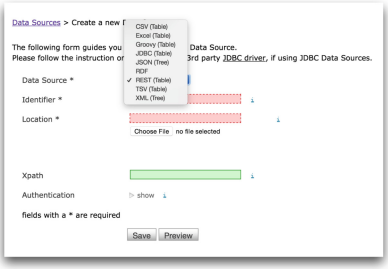
For some of its features, DataOps makes use of established external components that are integrated in the backend. In particular, ETL extraction of relational databases currently relies on DB2Triples[2] and entity reconcilation uses Silk [4]. All modules are plug-able due to generic interfaces and standards. Post-processor modules can even be stacked as pipelines of sub components. For example, other standard R2RML mapping engines can be hooked in, if required. The interface used for editing relational-to-ontology mappings is a newer version of the prototype presented in [5].

DataOps builds on the technology of a semantic platform, Information Workbench [6]. Information Workbench is a platform for Linked Data application development, including flexible and extensible interfaces for semantic integration, visualization and collaboration. It provides features such as managed triple store connectivity, SPARQL federation [7], ontology management, or exploration and visualization of resulting data. More specific features for data integration and business analytics use cases can be installed as Apps (e.g., more advanced visualization components, or automatic mapping support [8]).

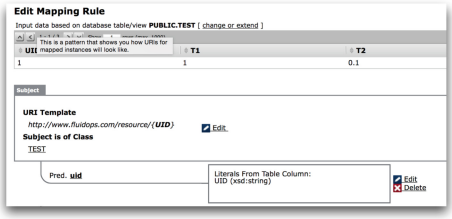
DataOps is available both under a commercial license and under Open Source terms, as part of the Information Workbench.[3]

Visitors of the demo will be able to perform all three steps of the integration process. In addition, visitors will be able to view and explore integrated data with widget provided by the platform.

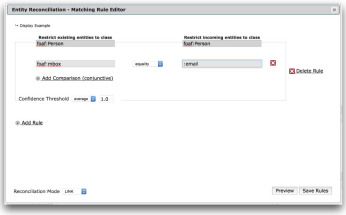
[2] https://github.com/antidot/db2triples/.
[3] http://www.fluidops.com/en/company/training/open_source.



(a) Step 1: Configuration of Data Sources



(b) Step 2: Example of a Mapping Editor (R2RML)



(c) Step 3: Reconcilation Rules

Fig. 2. Setup/Configuration steps demonstrated

3 Related Work

Individual connector, extractor, translation and editor components that can be hooked together for disparate ETL style tasks are commonly available, including some of those used in DataOps (e.g., [4–6]). For end-to-end solutions, however, they are often hard to configure and to use, especially when disparate data sources (e.g., relational databases, XML, Excel, and RDF) each require different components in the setup.

There are recent approaches to target this heterogeneity challenge in semantic data integration, such as RML [9], a unified mapping language to translate different source formats into RDF. However, those (and similar) approaches are still isolated components when it comes to providing users with an end-to-end toolkit.

A few end-to-end semantic data integration platforms exist so far. However, they usually do not focus on the needs of enterprise data integration but rather on other use cases (e.g., on RDF data sources or on Web scale data rather than enterprise data). For instance, UnifiedViews [10] provides easy to use interfaces, however, focuses on processing of pre-existing RDF data. LDIF [3] provides an interface for some of the surrounding tasks but also has a focus on RDF data, even though relational data is also supported.

Some recent projects (e.g., the Optique platform [11]) aim at providing an integrated platform for large-scale enterprise data integration but at the same time focus to push on other limitations such as federated, virtualized scale-out or even streaming data.

4 Conclusion

We presented DataOps, a seamless Anything-to-RDF semantic data integration toolkit. In contrast to most existing frameworks, DataOps assumes data sources in any number of mostly non-semantic source data formats, available from different locations within one organization. DataOps also integrates all procedural and lifecycle aspects of setting up and providing integrated data in a single platform, offering a seamless user experience. It is thus novel as a targeted solution for semantic enterprise data integration with a focus on many different, heterogeneous data formats and ease of use.

In the demo, users can try all aspects of DataOps for themselves using both locally provided data and data accessible from the Web.

References

1. Vavilapalli, V.K., et al.: Apache Hadoop YARN: yet another resource negotiator. In: SOCC 2013 (2013)
2. Evelson, B., Kisker, H., Bennett, M., Christakis, S.: Benchmark your BI environment. Technical report, Forrester Research, Inc., October 2013
3. Schultz, A., Matteini, A., Isele, R., Bizer, C., Becker, C.: LDIF - linked data integration framework. In: COLD 2011 (2011)
4. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009)
5. Pinkel, C., Binnig, C., Haase, P., Martin, C., Sengupta, K., Trame, J.: How to best find a partner? An evaluation of editing approaches to construct R2RML mappings. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 675–690. Springer, Heidelberg (2014)
6. Haase, P., Schmidt, M., Schwarte, A.: The information workbench as a self-service platform for linked data applications. In: COLD (2011)
7. Schwarte, A., Haase, P., Hose, K., Schenkel, R., Schmidt, M.: FedX: optimization techniques for federated query processing on linked data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 601–616. Springer, Heidelberg (2011)
8. Pinkel, C., Binnig, C., Kharlamov, E., Haase, P.: IncMap: pay-as-you-go matching of relational schemata to OWL ontologies. In: OM (2013)
9. Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., Walle, R.V.D.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: LDOW (2014)
10. Knap, T., Kukhar, M., Macháč, B., Škoda, P., Tomeš, J., Vojt, J.: UnifiedViews: an ETL framework for sustainable RDF data processing. In: ESWC (Posters & Demos) (2014)
11. Kharlamov, E. et al.: Optique 1.0: semantic access to big data - the case of Norwegian petroleum directorate factpages. In: ISWC (Posters & Demos) (2013)