A Semantic Integration Approach for Building Knowledge Graphs On-Demand

Diego Collarana[1,2](✉)

[1] Enterprise Information Systems (EIS), University of Bonn, Bonn, Germany
collaran@cs.uni-bonn.de
[2] Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS),
Sankt Augustin, Germany

Abstract. Information about the same entity may be spread across several Web data sources, e.g., people on the social networks (Social Web), product descriptions on e-commerce sites (Deep Web) or in public Knowledge Graphs (Web of data). The problem of integrating entities from heterogeneous Web data sources on-demand is still a challenge. Existing approaches propose expensive Extraction Transformation Loading (ETL) processes and rely on syntactic comparison of entity properties, leaving aside the semantics encoded in the data. We devise *FuhSen*, an integration approach that exploits search capabilities of Web data sources and semantics encoded in the data. *FuhSen* generates Knowledge Graphs in response to keyword-based queries. Resulting Knowledge Graphs describe the semantics of the integrated entities, as well as the relationships among these entities. *FuhSen* approach utilizes an ontology to describe the Web data sources in terms of content and search capabilities, and exploits this knowledge to select the sources relevant for answering a keyword-based query on-demand. The results of various empirical studies of the effectiveness of *FuhSen* suggest that the proposed integration technique is able to accurately integrate data from heterogeneous Web data sources into a Knowledge Graph.

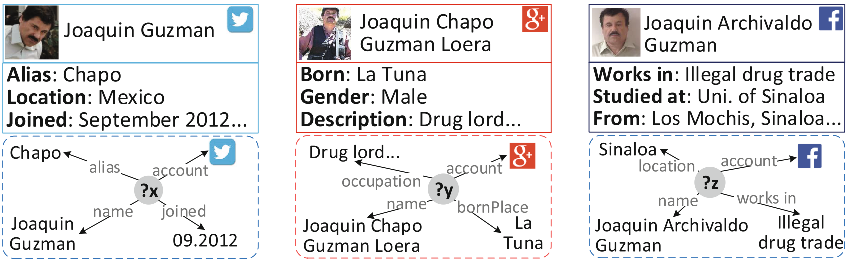
1 Problem Statement

The strong support that Web based technologies have received from researchers, developers, and practitioners has resulted in the publication of data from almost any domain on the Web. Additionally, standards and technologies have been defined to query, search, and manage Web accessible data sources. For example, Web access interfaces or APIs allow for querying and searching sources like Twitter, Google+, or the DBpedia, Wikidata. Web data sources make overlapping as well as complementary data available about entities, e.g., people or products. Although these entities may be described in terms of different vocabularies by these Web data sources, the data correspond to the same real-world entities. Thus, the distributed data needs to be integrated in order to have a more complete description of these entities (Fig. 1).

As example, consider a *distributed* and *heterogeneous* search scenario in the context of crime investigation. During a crime investigation process, collecting,

J. Cabot et al. (Eds.): ICWE 2017, LNCS 10360, pp. 575–583, 2017.
DOI: 10.1007/978-3-319-60131-1_45

and analyzing information from different sources is a key step performed by investigators. Although scene analysis is always required, a crime investigation process greatly benefits from searching information about people, products, and organisations on the Web. Typically, data collected from the following data sources is utilised for enhancing crime analysis processes: (1) The *Social Web* encompasses user generated content and personal profiles. (2) The *Deep Web* advertises products and services offered by organisations, e.g., the eBay e-commerce platform. (3) The *Web of Data* includes billions of machine-comprehensible facts, which can serve as background knowledge for collecting information about different types of entities. (4) The *Dark Web* refers to sites accessible only with specific software, and restricted trading of goods that can be accessed through the so-called dark-net markets.


Joaquin Guzman
Alias: Chapo
Location: Mexico
Joined: September 2012...
Chapo
alias
account
?x
name
joined
Joaquin Guzman
09.2012
Joaquin Chapo Guzman Loera
Born: La Tuna
Gender: Male
Description: Drug lord...
Drug lord...
occupation
account
?y
name
bornPlace
Joaquin Chapo Guzman Loera
La Tuna
Joaquin Archivaldo Guzman
Works in: Illegal drug trade
Studied at: Uni. of Sinaloa
From: Los Mochis, Sinaloa...
Sinaloa
location
account
?z
name
works in
Joaquin Archivaldo Guzman
Illegal drug trade

(a) Twitter data (b) Google+ data (c) Facebook data

Fig. 1. Motivating Example. Pieces of data (RDF molecules) about *Joaquin Chapo Guzman* collected from different Web social networks.

To solve this data integration scenario, in this doctoral work we propose *FuhSen*, a semantic integration approach that exploits Web APIs (e.g., REST APIs) provided by Web data sources to collect, and integrate molecules of data, to then enrich and summarize information about an entity (e.g., a suspect). Using Linked Data as the core technology, the objectives of *FuhSen* approach is to provide a novel integration technique able to: (1) integrate heterogeneous data extracted from APIs into a unified data schema on-demand; (2) create a Knowledge Graph on-demand with the data extracted from the different data sources; and (3) enrich this Knowledge Graph using semantic algorithms e.g., entity disambiguation, typing and entity summarization, and ranking.

2 Research Objectives

This doctoral work attempts to answer the following research questions:

> **RQ1**: Can Knowledge Graphs be populated with data collected from heterogeneous Web data sources on-demand?

To answer **RQ1**, we plan to explore and evaluate the use of RDF vocabularies to facilitate source selection and data fusion tasks.

RQ2: Can semantics encoded in RDF graphs be exploited to integrate data collected from heterogeneous data sources?

To answer **RQ2**, we will analyse how to use both the explicit semantics, e.g., Properties, Relations, and Hierarchy of classes, and also the implicit semantics.

RQ3: Can semantic similarity measures be able to enhance accuracy of the integration of data collected from heterogeneous data sources?

To answer **RQ3**, we will evaluate semantic similarity approaches that can be used in the context of data integration. Propose a new semantic similarity metric with the goal of data integration is also an option to answer RQ3.

3 State-of-the-art

Traditional approaches toward constructing **Knowledge Graphs (KG)**, e.g., NOUS [3], Knowledge Vault [7], or DeepDive [12], imply materialization of the designed graph built from (un-, semi-)structured sources. Therefore, a heavy Extraction Transformation Loading (ETL) process needs to be executed to integrate the data. In comparison, the novelty of *FuhSen* approach resides in a non-materialized Knowledge Graph and usage of semantics encoded in the data. Non-materialization supports efficient knowledge delivery on-demand. Further, *FuhSen* creates RDF molecules that unify and encode hybrid knowledge from heterogeneous sources in an abstract entity. Moreover, the problem of integrating RDF graphs is in the research focus for many years. Knoblock et al. [10] propose KARMA, a framework for integrating structured data sources. Schultz et al. [15] describe a Linked Data Integration Framework (LDIF) that provides a set of independent tools to support the process of interlinking RDF datasets. For instance, SILK [9] identifies `owl:sameAs` links among entities of two datasets and Sieve [11] performs data fusion. Although the aforementioned approaches are fast and effective, they require domain knowledge and significant manual effort while configuring the pipeline. In contrast, *FuhSen* is a universal **black box** technique that requires only a small number of high-level parameters, while enables users to adjust the system according to the application domain.

4 Proposed Approach

Given a keyword query, *FuhSen* executes the query over the relevant sources, and utilizes semantic similarity measures to determine the relatedness among the entities to be integrated. *FuhSen* creates a Knowledge Graph with the integrated entities at query time. A Knowledge Graph is composed of a set of entities, their properties, and relations among these entities. The Semantic Web technology stack provides the pieces required to define and build a Knowledge Graph. To properly understand these concepts, we follow the notation proposed by Arenas et al. [1], Piro et al. [13], and Fernandez et al. [8] to define RDF triples, Knowledge Graphs, and RDF molecules, respectively.

Definition 1 (RDF triple [1]). *Let* $\mathbf{I}$, $\mathbf{B}$, $\mathbf{L}$ *be disjoint infinite sets of URIs, blank nodes, and literals, respectively. A tuple* $(s, p, o) \in (\mathbf{I} \cup \mathbf{B}) \times \mathbf{I} \times (\mathbf{I} \cup \mathbf{B} \cup \mathbf{L})$ *is denominated an RDF triple, where* s *is called the subject,* p *the predicate, and* o *the object.*

Definition 2 (Knowledge Graph [13]). *Given a set* T *of RDF triples, a Knowledge Graph is a pair* $G = (V, E)$*, where* $V = \{s \mid (s, p, o) \in T\} \cup \{o \mid (s, p, o) \in T\}$ *and* $E = \{(s, p, o) \in T\}$.

Definition 3 (RDF Subject Molecule [8]). *Given an RDF graph* G*, an RDF subject-molecule* $M \subseteq G$ *is a set of triples* $t_1, t_2, \dots, t_n$ *in which* $subject(t_1) = subject(t_2) = \cdots = subject(t_n)$.

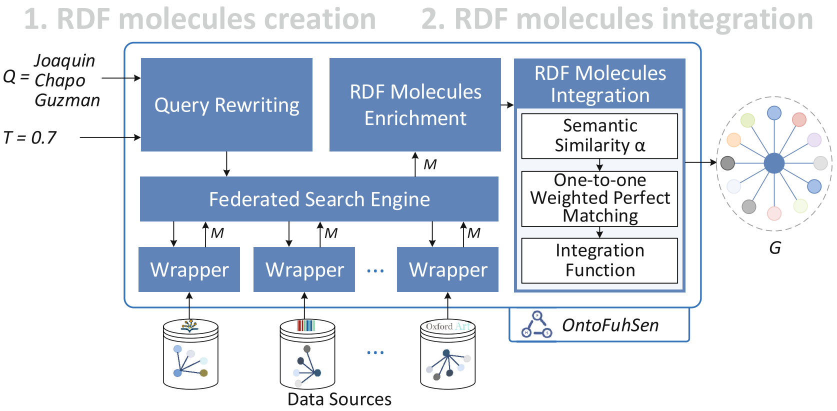


Fig. 2. The *FuhSen* Architecture. *FuhSen* receives a keyword query Q and a threshold T, and produces a Knowledge Graph G populated with the entities associated with the keywords

FuhSen is a two-fold approach, the architecture is shown in Fig. 2. The first step is the creation of RDF molecules of data from the heterogeneous Web data sources. Web services facilitate accessibility of data on-demand e.g., using REST services. In this step, RDF molecules from the same entities have to be recognized and integrated in order to build complete Knowledge Graphs

4.1 Creation of RDF Molecules

As an input, *FuhSen* receives a keyword query Q, e.g., *Joaquin Chapo Guzman*, and a similarity threshold value T, e.g., 0.7. The input values are processed by the *Query Rewriting* module, which formulates a correct query to be sent to the *Search Engine* module. The *Search Engine* explores several wrappers and transforms the output into RDF molecules. Intermediate results are enriched with additional knowledge in the *RDF Molecules Enrichment* module.

4.2 Integration of RDF Molecules

This module constructs a Knowledge Graph out of the enriched molecules. The input is a set of RDF molecules, and the output is an integrated RDF graph. The module consists of three sub-modules:

- *Computing Similarity of RDF Molecules.* Similar RDF molecules should be integrated in order to create a fused, universal representation of a certain entity. In contrast with triple-based linking engines like Silk [15], we employ a RDF molecule-based approach increasing the complexity level and considering the semantics of molecules. That is, we do not work with independent triples, but rather with a set of triples belonging to a certain subject. The RDF molecule-based approach allows for natural clustering of a Knowledge Graph, reducing the complexity of the linking algorithm.
- *1-1 Weighted Perfect Matching.* Given a weighted bipartite graph BG of RDF molecules, where weights correspond to values of semantic similarity between the RDF molecules in BG, a matching of BG corresponds to a set of edges that do not share an RDF molecule, and where each RDF molecule of BG is incident to exactly one edge of the matching. The problem of the 1-1 weighted perfect matching of BG corresponds to a matching where the sum of the values of edge weights in the matching have a maximal value.
- *Integration functions.* When similar molecules are identified under the desired conditions, the last step of the pipeline is to integrate them into an RDF Knowledge Graph. The result Knowledge Graph contains all the unique facts of the analyzed set of RDF molecules. The implementation of the integration function in *FuhSen* is the union, i.e., the logical disjunction, of the molecules identified as similar during the previous steps.

5 Research Methodology and Research Design

The research methodology of this doctoral work includes the following steps:

1. Review the literature to evaluate the state-of-the-art approaches relevant to the problem of integrating heterogeneous Web data sources on-demand.
2. Formalise an on-demand semantic integration approach named *FuhSen.*
3. Empirically evaluate different properties of the approach, e.g., effectiveness and performance. Evaluate different components of the architecture and propose new algorithms and operators to realize the vision of this work.

6 Results and Contributions

So far, we have evaluated the architecture and the effectiveness of *FuhSen*. In [4,6], we proposed and implemented an RDF vocabulary mediator-wrapper architecture and proposed an evaluation study to answer **RQ1**.

Lessons Learned: The proposed architecture implemented in *FuhSen* is able to query heterogeneous Web data sources and create RDF molecules of data in a federated manner. The results of our evaluations shows that the vocabulary approach defined in *FuhSen* architecture allows for handling heterogeneity of data in an effective way. At the same time more Web data sources can be plug in/out in an easy manner, as consequence, the integration process is reduced. However, the experiments show problems in terms of scalability, the more Web data sources the slower the integration process becomes. Thus, a better resource selection approach should be investigated to answer research question **RQ1**.

In [5], we propose a two-fold approach to integrate RDF molecules from different data sources, with this approach and its evaluation we answer research questions **RQ2** and **RQ3**. We evaluate the effectiveness of integration on different datasets. We also experiment with two similarity metrics: Jaccard and GADES [14]. Our goal was to determine the impact of similarity function on the integration approach. Therefore, a triplet-based similarity metric Jaccard is compared against a semantic similarity function GADES [14].

We created a **Gold Standard (GS)** of the type Person extracted from DBpedia, which results in 829,184 triples. Two **Test data Sets (TS)** were created from the Gold Standard with their properties and values randomly split among two test datasets. Each triple is randomly assigned to one or several test datasets. We measure the behavior of our integration approach *FuhSen* [5] in terms of the following metrics: *Precision*, *Recall*, and *F-measure* during the experiments. Precision is the fraction of RDF molecules that has been identified and integrated by the approach (M) that intersects with the Gold Standard (GS), i.e., $Precision = \frac{|M \cap GS|}{|M|}$. Recall corresponds to the fraction of the identified similar molecules in the Gold Standard, i.e., $Recall = \frac{|M \cap GS|}{|GS|}$.

Lessons Learned: Table 1 shows the effectiveness of *FuhSen* on the integration task over 20,000 molecules. Jaccard demonstrates lower performance on the data set as its algorithm just relies on the particular properties of the RDF molecule. Jaccard does not utilize semantics encoded in the Knowledge Graph. On the other hand, GADES exhibit a good performance and it might be used as a **black box** in *FuhSen* approach.

Performance of the integration depends on the threshold parameter. As a simple set-based approach, the performance (precision, recall, and F-measure) of the Jaccard similarity quickly decreases with higher thresholds, while GADES remains stable. These insights suggest a positive answer to research questions **RQ2** and **RQ3**. However, GADES performance is impacted by the quality of the schema e.g., a good design of the hierarchy of classes, properties, and relations. Thus, enrichment of the molecules and tuning process is a pre-requisite in GADES. This impacts on an automatic nature of the problem we are trying to solve. Therefore, a pre-trained and automatic similarity function to compare RDF Molecules is required to answer research questions **RQ2** and **RQ3**.

Table 1. Effectiveness of *FuhSen* on 20,000 RDF molecules. Jaccard vs GADES approach using different thresholds (T). Highest values of Recall and F-measure are highlighted in bold.

	T0.0	T0.1	T0.2	T0.3	T0.4	T0.5	T0.6	T0.7	T0.8	T0.9
Precision										
Jaccard	0.72	0.77	0.44	0.34	0.37	0.36	0.27	0.21	0.21	0.21
GADES	0.76	**0.80**	**0.80**	0.79	0.79	0.79	0.79	0.76	0.70	0.65
	T0.0	T0.1	T0.2	T0.3	T0.4	T0.5	T0.6	T0.7	T0.8	T0.9
Recall										
Jaccard	0.72	0.42	0.09	0.05	0.02	0.02	0.01	0.01	0.01	0.01
GADES	**0.76**	**0.76**	**0.76**	**0.76**	**0.76**	**0.76**	0.68	0.46	0.22	0.06
	T0.0	T0.1	T0.2	T0.3	T0.4	T0.5	T0.6	T0.7	T0.8	T0.9
F-Measure										
Jaccard	0.72	0.54	0.15	0.08	0.04	0.04	0.02	0.02	0.02	0.02
GADES	0.76	**0.78**	**0.78**	0.77	0.77	0.77	0.73	0.57	0.33	0.11

Although significant progress has been done in the context of this doctoral work, more empirically results are needed to fully answer research questions **RQ1**, **RQ2**, and **RQ3**. Next section describes the plan for the next year.

7 Work Plan

This doctoral work is entering in its final stage (3rd year). To completely answer the defined research questions the following research tasks remain:

1. Propose a novel RDF fusion operator, the operator should be able to determine the relatedness between two RDF molecules and integrating them. This task is related to **RQ2**, and the target publication is a research paper.
2. Present a semantic similarity measure based on TransE [2] which utilizes the gradient descent optimization method to learn the features representation of RDF entities automatically. This task is related to **RQ3**, and the target publication is a research paper.
3. Present a scalable and efficient source selection based on the semantic description of the Web sources and keyword query. This task is related to **RQ3**, and we plan to publish a research paper.

8 Conclusions

In this doctoral work, we address the problem of data integration about the same entity that is spread in different Web data sources. We propose *FuhSen*, an *on-demand semantic integration approach* that creates Knowledge Graphs on-demand by integrating data collected from a federation of heterogeneous data sources using an RDF molecule integration approach. We have explained the creation of RDF molecules by using Linked Data wrappers; we have also presented how semantic similarity measures can be used to determine the relatedness of two resources in terms of the relatedness of their RDF molecules. Results of the empirical evaluation suggest that *FuhSen* is able to effectively integrate pieces of information spread across different data sources. The experiments suggest that the molecule based integration technique implemented in *FuhSen* integrates data in a Knowledge Graph more accurately than existing integration techniques.

Acknowledgments. I would like to thank Maria-Esther Vidal, Christoph Lange, Sören Auer, and the German Ministry of Education and Research with grant no. 13N13627 project LiDaKrA for supporting this work.

References

1. Arenas, M., Gutierrez, C., Pérez, J.: Foundations of RDF databases. In: Tessaris, S., Franconi, E., Eiter, T., Gutierrez, C., Handschuh, S., Rousset, M.-C., Schmidt, R.A. (eds.) Reasoning Web 2009. LNCS, vol. 5689, pp. 158–204. Springer, Heidelberg (2009). doi:10.1007/978-3-642-03754-2_4
2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems 26. Curran Associates Inc. (2013)
3. Choudhury, S., Agarwal, K., Purohit, S., Zhang, B., Pirrung, M., Smith, W., Thomas, M.: Nous: construction and querying of dynamic knowledge graphs. arXiv preprint arXiv:1606.02314 (2016)
4. Collarana, D., Galkin, M., Lange, C., Grangel-González, I., Vidal, M., Auer, S.: Fuhsen: a federated hybrid search engine for building a knowledge graph on-demand (short paper). In: OTM Conferences - ODBASE (2016)
5. Collarana, D., Galkin, M., Traverso-Ribón, I., Lange, C., Vidal, M.-E., Auer, S.: Semantic data integration for knowledge graph construction at query time. In: ICSC (2017)
6. Collarana, D., Lange, C., Auer, S.: Fuhsen: a platform for federated, RDF-based hybrid search. In: WWW Companion Volume (2016)
7. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: SIGKDD. ACM (2014)
8. Fernández, J.D., Llaves, A., Corcho, O.: Efficient RDF Interchange (ERI) format for RDF data streams. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014. LNCS, vol. 8797, pp. 244–259. Springer, Cham (2014). doi:10.1007/978-3-319-11915-1_16

9. Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. J. Web Seman. **23**, 2–15 (2013)
10. Knoblock, C.A., et al.: Semi-automatically mapping structured sources into the semantic web. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 375–390. Springer, Heidelberg (2012). doi:10.1007/978-3-642-30284-8_32
11. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, 30 March 2012 (2012)
12. Palomares, T., Ahres, Y., Kangaspunta, J., Ré, C.: Wikipedia knowledge graph with DeepDive. In: 10th AAAI Conference on Web and Social Media (2016)
13. Pirrò, G.: Explaining and suggesting relatedness in knowledge graphs. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9366, pp. 622–639. Springer, Cham (2015). doi:10.1007/978-3-319-25007-6_36
14. Ribón, I.T., Vidal, M., Kämpgen, B., Sure-Vetter, Y.: GADES: a graph-based semantic similarity measure. In: SEMANTiCS (2016)
15. Schultz, A., Matteini, A., Isele, R., Mendes, P.N., Bizer, C., Becker, C.: LDIF - a framework for large-scale linked data integration. In: 21st International World Wide Web Conference, Developers Track, April 2012