Datenbank Spektrum
https://doi.org/10.1007/s13222-019-00312-z

FACHBEITRAG

# Transforming Heterogeneous Data into Knowledge for Personalized Treatments—A Use Case

**Maria-Esther Vidal[1] · Kemele M. Endris[1] · Samaneh Jazashoori[1] · Ahmad Sakor[1] · Ariam Rivas[1]**

Received: 27 February 2019 / Accepted: 30 March 2019


**Abstract**
Big data has exponentially grown in the last decade; it is expected to grow at a faster rate in the next years as a result of the advances in the technologies for data generation and ingestion. For instance, in the biomedical domain, a wide variety of methods are available for data ingestion, e.g., liquid biopsies and medical imaging, and the collected data can be represented using myriad formats, e.g., FASTQ and Nifti. In order to extract and manage valuable knowledge and insights from big data, the problem of data integration from structured and unstructured data needs to be effectively solved. In this paper, we devise a knowledge-driven approach able to transform disparate data into knowledge from which actions can be taken. The proposed framework resorts to computational extraction methods for mining knowledge from data sources, e.g., clinical notes, images, or scientific publications. Moreover, controlled vocabularies are utilized to annotate entities and a unified schema describes the meaning of these entities in a *knowledge graph*; entity linking methods discover links to existing knowledge graphs, e.g., DBpedia and Bio2RDF. A federated query engine enables the exploration of the linked knowledge graphs while knowledge discovery methods allow for uncovering patterns in the knowledge graphs. The proposed framework is used in the context of the EU H2020 funded project iASiS with the aim of paving the way for accurate diagnostics and personalized treatments.

## 1 Introduction

Integrating data-driven digital technologies in conjunction with smart infrastructures for management and analytics, increasingly, offer huge opportunities for improving quality of life [40] and industrial competitiveness [56]. However, the enormous amount of data generated in scientific and industrial domains, demands the development of computational methods for ingestion, integration, and analysis, as well as for the transformation of big data into knowledge. The problem of data integration has been extensively addressed by the Database community [15, 22]. As a result, a vast amount of integration frameworks [11, 21, 28, 31, 37] have been developed; they implement data integration systems following the local-as-view (LAV) and global-as-view (GAV) paradigms[32]. Further, query processing has also played a relevant role in solving data integration on the fly. Graph-based traversal [33, 7], and federated query processing [1, 16] are representative approaches for enabling data integration at query execution time. Although these approaches have made remarkable contributions, the problem of scaling up to big data transformation remains unsolved. The lack of techniques able to manage structured and unstructured sources (e.g., clinical notes, images, scientific publications) is the main drawback of existing approaches.

**Our Research Goal:** Our main objective is to tackle data integration of structured and unstructured data in a way that the meaning of integrated data can be described, explored, and used to uncover relevant insights. We focus on biomedical data sources in the context of the EU H2020 funded project iASiS[1] and show how the problem of data integration may hinter the prescription of personalized treatments. Given a collection of data sets (structured and unstructured), the problem of data integration is to identify if two entities in the data sets match or do not match the same real-world entity. Integrating data sets requires the recognition and resolution of interoperability conflicts across these data sets, as well as fusion policies for merging equivalent entities [10]. Considering the wide nature of entities, the state of the art has focused on integration methods that reduce manual work and maximize accuracy and precision [11, 18, 19]. To overcome interoperability conflicts generated by the wide

✉ Maria-Esther Vidal
Maria.Vidal@tib.eu

[1] TIB Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167 Hannover, Germany

[1] http://project-iasis.eu/.

Published online: 15 April 2019

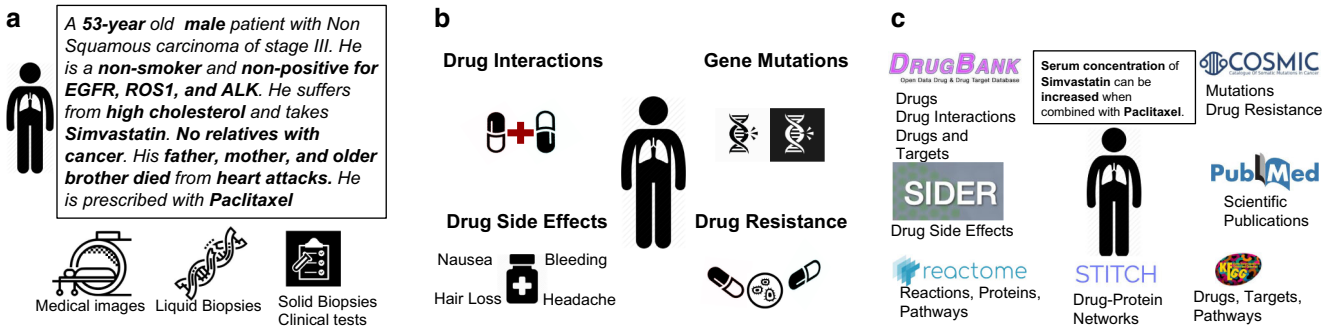


**Fig. 1** **Motivating Example**. Heterogeneous sources of knowledge. **a** Unstructured data sources, e.g., clinical notes, medical images, and clinical tests, encode invaluable knowledge about a patient medical condition. **b** Factors impact on the effectiveness of a treatment; they need to be identified to increase a patient survival time. **c** Various biomedical repositories maintain knowledge collected by the scientific community about facts that can contribute to the prescription of effective treatments. Data sources range from structured (e.g, COSMIC), to unstructured (e.g., PubMed); and short texts in structured data sources may encode also relevant knowledge (e.g., drug interactions). Heterogeneity problems across sources need to be solved for extracting the required knowledge. **a** Electronic health records, **b** impacts in treatment effectiveness, **c** biomedical data sources

variety of existing formats–short notes or scientific publications–several unstructured processing techniques have been proposed. Natural language processing (NLP) contributes to integrating structured and textual data by providing linguistic annotation methods at different levels [36, 38, 41], e.g., syntactic parsing, named entity recognition, word sense disambiguation, and entity linking. Further, visual analytics techniques facilitate the extraction and annotation of entities from non-textual data sources [27, 5]. Annotations from ontologies and controlled vocabularies extracted from unstructured data represent the basis for determining relatedness among the annotated entities by the mean of similarity measures, as well as for identifying matches between highly similar entities.

**Approach:** The main idea of this paper is to present a knowledge-driven framework that resort to knowledge extraction, ontologies, and data integration techniques in order to create a knowledge graph. It comprises data and the knowledge that describes the main characteristics of the integrated data. The proposed approach represent a building block for the support of clinicians during disease diagnosis and treatment prescription.

**Contributions:** The principal contributions of this paper are the presentation of the results of applying the knowledge-driven framework to various biomedical data sources, as well as the promising outcomes observed by analyzing the generated knowledge graph. Although the framework as a whole is not available as open source, the components to perform entity linking[2] and knowledge graph management[3] are publicly available. The remainder of this article is structured as follows: Sect. 2 motivates the data integration problem over biomedical data sets. Sect. 3 describes our knowledge-driven framework, and Sect. 4 summarizes the principal results of implementing this framework in the iASiS project. Related work is presented in Sect. 5, and finally, Sect. 6 concludes and give insights for future work.

## 2 Motivating Example

We motivate our work with a set of myriad sources of knowledge about the condition of a lung cancer patient (Fig. 1), as well as typical integration problems caused as a result of well-known data complexity issues, e.g., variety, volume, and veracity. Electronic health records (EHRs) (Fig. 1a) preserve the knowledge about the conditions of a patient that need to be considered in order for effective diagnoses and treatment prescriptions. Albeit informative, EHRs usually preserve patient information in an unstructured way, e.g., textual notes, images, or genome sequencing. Furthermore, EHRs may include incomplete and ambiguous statements about the whole medical history of a patient. In consequence, knowledge extraction techniques are required to mine and curate relevant information for an integral analysis of a patient, e.g., age, gender, life habits, mutations, diagnostics, treatments, and familial antecedents. In addition to evaluating information in EHRs, physicians depend on their experience or available sources of knowledge to predict potential adverse outcomes, e.g., drug interactions, side-effects or resistance (Fig. 1b). Diverse repositories and databases make available crucial knowledge for the complete description of a patient condition and the potential outcome (Fig. 1c). Nevertheless, sources are autonomous and utilize diverse formats that range from unstructured scientific publications in PubMed[4] to dumps of structured data about cancer related mutations in COSMIC[5]. To illustrate,

[2] https://labs.tib.eu/info/en/project/falcon/.

[3] https://github.com/SDM-TIB/KG-Tools.

[4] https://www.ncbi.nlm.nih.gov/pubmed/.

[5] https://cancer.sanger.ac.uk/cosmic.

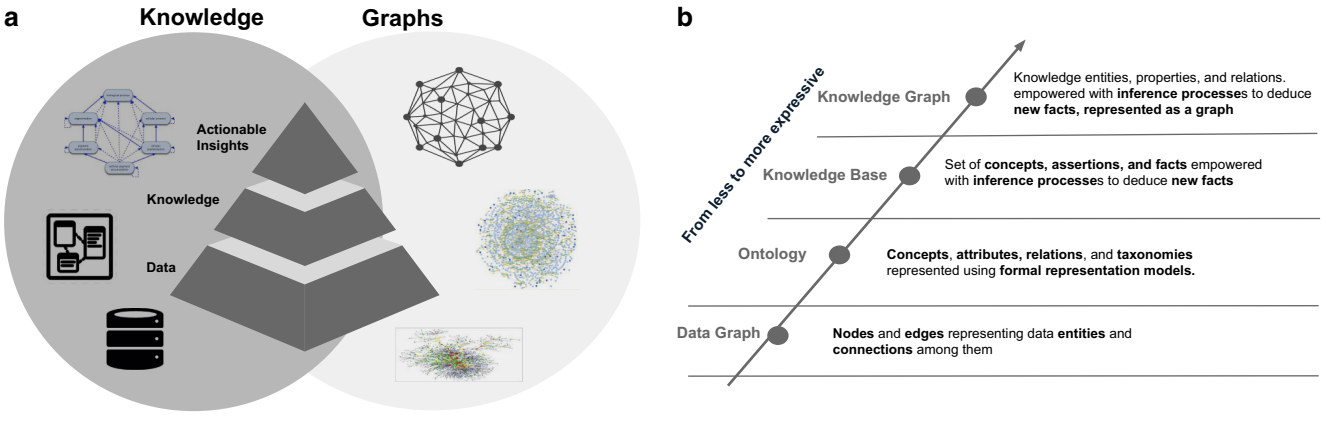


**Fig. 2** **Definition of a Knowledge Graph**. **a** A knowledge graph is presented as the intersection of the formal models able to represent facts of various types and levels of abstraction using a graph-based formalism. **b** Knowledge representation models are characterized according to the represented facts and levels of abstraction. **a** Knowledge graph, **b** a spectrum of knowledge representation

the effect of the interactions between two drugs is reported in DrugBank like short text, e.g., the effect of the interactions between Simvastatin and Paclitaxel. In order to detect the facts that can impact on the effectiveness of a particular treatment, e.g., Paclitaxel, the physician will have to search through these diverse data sources and identify the potential adverse events and interactions. Data complexity issues like data volume and diversity impede an efficient integration of the knowledge required to predict the outcomes of a treatment.

The proposed knowledge-driven framework resorts to techniques of knowledge extraction and representation to create a knowledge graph where data from disparate data sources is integrated. A knowledge graph represents entities and their relations, and ontologies and controlled vocabularies are utilized to describe the meaning of relations, as well as for annotating entities in a uniform way in the knowledge graph. Unified Medical Language System (UMLS), the Human Phenotype Ontology (HPO), and the Gene Ontology (GO) are exemplar ontologies. Furthermore, entity linking techniques are part of the framework to allow for the linking of entities in the knowledge graph, e.g., the drug Paclitaxel, to equivalent entities in existing knowledge graphs, e.g., in DBpedia[6] and in Bio2RDF[7]. The linked knowledge graphs composed a federation, and a federated query engine is able to execute queries against the various knowledge graphs. Finally, (un)supervised techniques are built on top of the knowledge graphs for the support of conscientious diagnosis and personalized treatments.

[6] http://dbpedia.org/resource/Paclitaxel.

[7] http://bio2rdf.org/drugbank:DB01229.

## 3 Our Approach

### 3.1 Preliminaries

Fig. 2 presents the main characteristics of a knowledge graph. First, a knowledge graph is depicted as a data structured that represents data, knowledge, and actionable insights using a graph data model (Fig. 2a). Graph data models enable for the representation of entities and their relations, as well as naturally model mono- and multi-valued attributes, the neighborhoods of an entity, different types of relations, and relations recursively specified. Moreover, graph data models naturally extend to a large number of relations between two entities and enable the traversal and exploration of these connections. Based on these features of graph models, they stand for suitable data models for representing different types of concepts in a knowledge graph. We define a knowledge graph as follows:

**Definition 1** A Knowledge Graph is a directed graph defined as triple $KG = (O, V, E)$, where:

- O is an ontology that comprises classes and relations, as well as rules that define the meaning of the relations.
- V is a set of nodes in the knowledge graph; nodes in V correspond to classes or instances of classes in O.
- E is a set of directed labeled edges in the knowledge graph that relate nodes in V. Edges are labeled with relations in O.

As stated in the previous definition, nodes in a knowledge graph can be composed of entities of representing items of data, abstract concepts, or the combination of both. This property enables for the characterization of a spectrum of knowledge graphs as indicated in Fig. 2b. This spectrum goes from less to more expressive graphs. *Data graphs* correspond to less expressive knowledge graphs; they com-

prise nodes representing entities and edges depicting the relations between them. Semantics of the relations is not encoded in any way in the graph. *Ontologies* include abstract concepts or classes—represented as nodes—and predicates representing the relations of these classes—edges in an ontology; the meaning of the predicates is represented using rules. *Knowledge bases* model knowledge about facts and abstract concepts but not necessarily using a graph data model; rule based formalisms like Datalog [8] or PSL [20] have been used to represent knowledge bases. Finally, *knowledge graphs* comprise not only facts about entities and their relations, but also about the classes to which these entities belong to and the meaning of these relations. Differently to knowledge bases, knowledge graphs are represented using graph data models; thus, they are able to naturally model data—entities—and knowledge—meaning of relations—as first-class citizens. Additionally, knowledge graphs can be modeled using diverse knowledge representation formalisms; the selection of the formalism depends on the type of statements that will be expressed in a knowledge graph. For example, the Resource Description Framework (RDF) is a metadata data model that resorts to the idea of making statements about resources in expressions of the form *subject-predicate-object*, known as triples. Subjects are represented as resources in the form of URIs or blank nodes; predicates define the relation between *subject and object*; they are in the form of URIs while *objects* can be of any type. RDF Schema is an extension of the basic RDF that allows for the definition of classes, relations, as well as hierarchies of classes and relations. Moreover, more expressive formalisms like the Ontology Web Language (OWL), make available a larger number of operators which enable the representation not only of classes, relations, and hierarchies, but also class and property constraints, negative statements, general equivalence relations, and restrictions of cardinality. In the knowledge graphs considered in this paper, operators from RDF, RDFS, and OWL are used. Further, some predicates have been also utilized to express metadata about classes and relations. For example, predicates rdfs:label, rdfs:comment, dcterms:modified, and dcterms:creator describe labels, comments, last modification date, and the creator of classes and properties, respectively. Data sources depicted in Fig. 1 are characterized by various conflicts that hinter a scalable solution of the problem of data integration. Heterogeneity conflicts include:

(i) **Structuredness** depends on the degree of the sources being structured.
(ii) **Schematic** is present whenever various schema are utilized by the data sources.
(iii) **Domain** occurs if different interpretations of the same universe of discourse are followed.
(iv) **Representation** takes place whenever different representations are used to model the same concept.
(v) **Language** exists among two or more data sources whenever different languages are utilized for modeling data or metadata.
(vi) **Granularity** depends on the graininess used to represent the data in different data sources.

### 3.2 A Knowledge-driven Framework

We devise a knowledge-driven framework able to transform and integrate heterogeneous data into knowledge graphs. Fig. 3 depicts an overview of the framework; it is composed of four main components: Data Ingestion, Semantic Data Integration, Exploration and Visualization, and Evaluation and Knowledge Discovery.

**1-Data Ingestion:** big data is collected from different data sources; collected data is mainly characterized by the three dominant dimensions of the Vs model: volume—very large data sets; variety—sources in multiple data formats and models; and veracity—data with potential biases, ambiguities, and noise. To overcome interoperability issues caused by data variety, distinct knowledge extraction methods are part of the framework. Typical extractions methods include:

(i) *Natural language processing* to extract facts from unstructured data sources and represent the extracted knowledge in the form of triples, i.e., subject, predicates, and objects [41]. Ontologies and controlled vocabularies are used to guide the extraction process as well as to annotate the extracted facts with the terms.
(ii) *Visual analysis and image processing* to extract relevant facts from non-textual material like videos and images [27, 5].
(iii) *Genomic Analysis* to identify mutations and genetic variations from microarrays [30, 53].

Once data is ingested, different techniques are used for data curation, e.g., statistical methods for completing missing values [46]–multiple imputation and maximum likelihood estimation, clustering techniques for duplicate detection [26], and crowdsourcing for data curation [4].

**2-Semantic Data Integration:** The integration of the matching entities is performed over the knowledge graph by exploring concepts, relations, taxonomies, and rules represented in the knowledge graph. First, collected and curated big data is modeled using a unified schema and stored in a knowledge graph. Then, entity recognition and linking are employed for transforming textual values in the knowledge graph, e.g., descriptions and comments, into structured facts. Finally, different methods are combined to curate and complete the represented facts. Knowledge graph creation

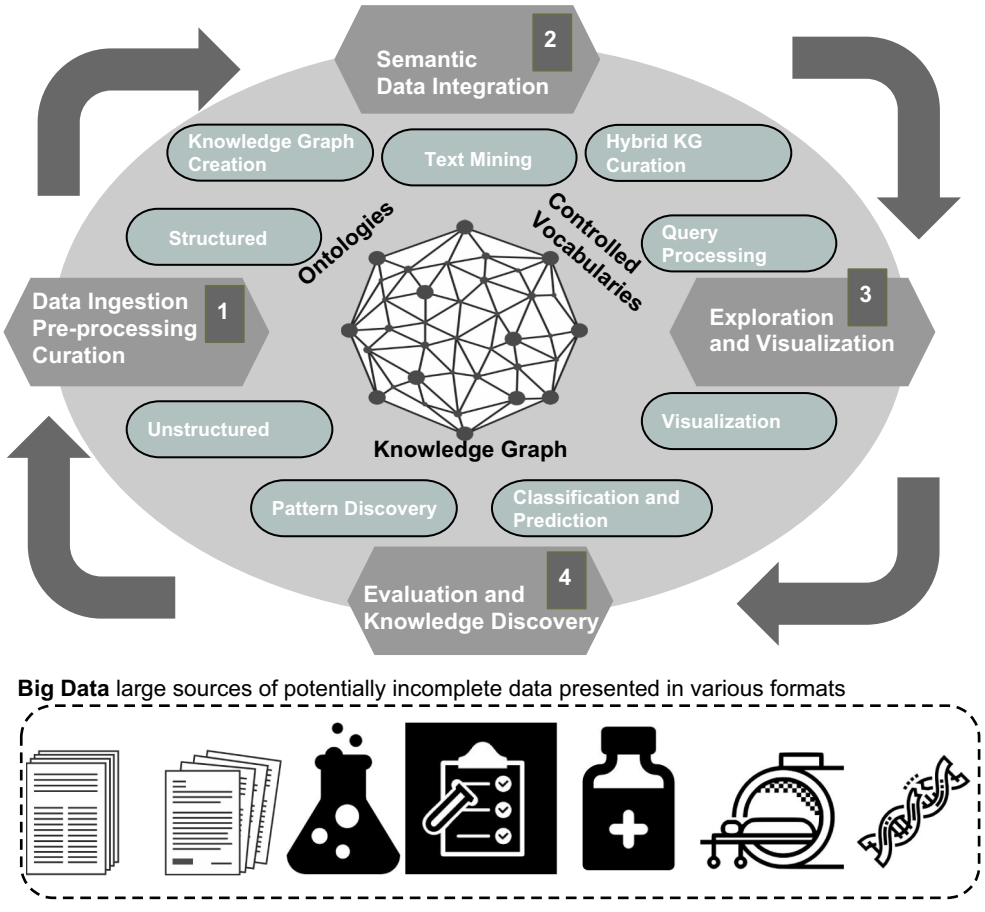


**Fig. 3 Knowledge Graph Overview**. Big data sources are ingested, curated, and integrated into a knowledge graph. Diverse knowledge extraction methods enable to transform unstructured data and describe the extracted facts using ontologies. Federated query processing and visualization tools enable the exploration of the knowledge graph, and knowledge discovery techniques facilitate the uncovering of relevant patterns

relies on mapping-driven algorithms guided by mapping rules that describe entities using a unified schema. Additionally, controlled vocabularies and ontologies used by the knowledge extraction tools are represented as RDF triples as well; links between these ontologies are also included in the knowledge graph to enable the identification of entities in different vocabularies. Similarity-based methods are performed for entity matching; similarity measures are able to exploit the knowledge encoded in the knowledge graph. Hybrid approaches combine reasoning processes on top of the knowledge graph with the wisdom of experts, and enable curation and knowledge completion [2].

**3-Exploration and Visualization:** SPARQL endpoints enable the independent access of knowledge graphs; they are Web services that provide Web interfaces to query RDF data following the SPARQL protocol. Queries against federations of SPARQL endpoints are posed through federated SPARQL query engines; they are devised following the generic mediator and wrapper architecture [54, 55]. Lightweight wrappers translate SPARQL subqueries into the required SPARQL endpoint calls and translate endpoint answers into the query engine internal structures. The mediator rewrites original queries into subqueries that can be executed by the SPARQL endpoints. Furthermore, the mediator gathers the results of evaluating subqueries, and combines the results to produce the answer of the query. The federated query engine is able to exploit the semantics encoded in the knowledge graph during the execution of the tasks of source selection, and query decomposition, optimization, and execution. Visualization tools facilitate the exploration of patterns in the knowledge graph.

**4-Evaluation and Knowledge Discovery:** Machine learning methods are utilized to identify patterns in the knowledge graph. These methods are enhanced with contextual knowledge represented in the knowledge graph with the aim of identifying accurate predictions whose meaning can be described.

## 4 The Knowledge-driven Framework for Supporting Personalized Medicine

iASiS is a 36-month H2020-RIA project that has started in April 2017. iASiS aims at transforming clinical and pharmacogenomics big data into actionable knowledge for the support of personalized medicine in two life-threatening diseases: lung cancer and dementia. The knowledge-driven framework depicted in Fig. 3 is applied to integrate

anonymized clinical data, biological sample analysis, medical images, genomics, medications, and scientific publications into the iASiS knowledge graph. UMLS and HPO are used for annotating concept extracted from unstructured items of data. the instantiation of the framework is as follows:

**1-Data Ingestion:** Knowledge extraction methods (developed by different partners) of the iASiS project are described as follows:

(i) *Electronic Health Record (EHR) Analysis:* NLP methods (Menasalvas et al. [36]) resort to named entity recognition to extract relevant entities from unstructured clinical notes and to annotate the extracted concepts with terms from UMLS. These techniques allow for the extraction of the 39 properties from 739 lung cancer patients [50].

(ii) *Genomic Analysis:* Data mining tools, e.g., catRapid (Livi et al. [34]), identify protein-RNA associations with high accuracy. Publicly available datasets, e.g., data from GTEx, GEO, and ArrayExpress, are used for the integration with transcriptomic data; genes are annotated with identifiers from different databases, e.g., HUGO or Uniprot/SwissProt, as well as with HPO.

(iii) *Image Analysis:* Several machine learning algorithms (Ortiz et al. [44]), are applied to learn predictive models able to classify medical images and detect areas of interests, e.g., lung cancer tumors or imaging biomarkers. Further, image annotation methods semantically describe these areas of interest using ontologies [12, 47]. *Open Data Analysis:* An NLP pipeline is followed to extract UMLS terms from scientific publication in PubMed[8] and relations between the extracted terms (Nentidis et al. [42]). This pipeline resorts to MetaMap[9] for UMLS term extraction and SemRep[10] for relation extraction. The NLP pipiline has enabled the collection of 166,073 UMLS terms from 250,688 publications.

**2-Semantic Data Integration.** Data collected from biomedical open data sources and the data sets generated from the knowledge extraction methods are integrated in this step. Open data sources include COSMIC[11], DrugBank[12], SIDER[13], and STITCH[14]. Albeit structured, the open data sets may contain unstructured fields that encode valuable knowledge, e.g., the description of the interactions between two drugs from DrugBank or the approved indications of a drug in DBpedia[15]. Entity and predicate linking methods (Sakor et al. [51]) are employed to extract entities and relations and to link them to terms in UMLS or DBpedia. A unified schema is used to represent the data in the iASiS knowledge graph. GAV mappings expressed using the RDF Mapping Language (RML) [14], specify mapping rules to transform data into RDF triples in the iASiS knowledge graph. Fig. 4a depicts with an example, the pipeline followed to create RDF triples and to perform data integration. EHR analysis [36] is performed to extract relevant facts from the clinical notes and represent these facts using UMLS. For simplicity, we just present some of facts: age, gender, toxic habits, chemotherapy drugs, drugs for comorbities, familial antecedents, and mutated genes (EGFR, ALK, ROS1). The execution of the RML mappings enable the creation of an RDF graph describing the patient and its relations. Note that drugs for chemotherapy and comorbities are annotated with the corresponding UMLS terms, i.e., C0144576 and C0074554 for Paclitaxel and Simvastatin, respectively. In addition, entity and predicate liking [51] is performed and effect of the interaction between Paclitaxel and Simvastatin is represented as an RDF graph (step 2). Since, this data is extracted from DrugBank, drugs are identified with a DrugBank identifier. Matching between the UMLS and DrugBank identifiers are found by performing string matching between the name of the drugs in DrugBank and the *preferred* names in UMLS. Matchings between UMLS and DrugBank identifiers—represented as dashed lines—are used for generating an RDF graph that relates UMLS identifiers of Paclitaxel and Simvastatin with the effects and impact of the interactions (step 3). Fig. 4 presents the final RDF graph where the patient described in the clinical notes and the interactions between his treatments are represented in an RDF graph. The same data integration procedure is performed for associating a patient with the side-effects of his/her prescribed drugs, the scientific publications in PubMed where his/her conditions, treatments, and biomarkers are reported; information about the diseases associated with his/her mutations; and potential mutations that may impact the effectiveness of his/her treatments. Moreover, the entity and predciate liking techniques by Sakor et al. [51] are also utilized to link entities in the iASiS knowledge graph with equivalent entities in DBpedia and Bio2RDF. Only for drugs, the approach by Sakor et al. [51] was able to identify 960 correct links to DBpedia out of 968 Drugs, while DBpedia Spotlight [13], a state-of-the-art entity linking tool, only identified 929 correct links.

The current version of the iASiS knowledge graph has 1,3 Billion triples, 46 RDF classes, in average 6.98 relations per entity, and each class is connected in average to 2.87 classes. Classes include Drugs, Publications, Muta-

[8] https://www.ncbi.nlm.nih.gov/pubmed/.

[9] https://metamap.nlm.nih.gov/.

[10] https://semrep.nlm.nih.gov/.

[11] https://cancer.sanger.ac.uk/cosmic.

[12] https://www.drugbank.ca/.

[13] http://sideeffects.embl.de/.

[14] http://stitch.embl.de/.

[15] http://dbpedia.org/page/Paclitaxel.

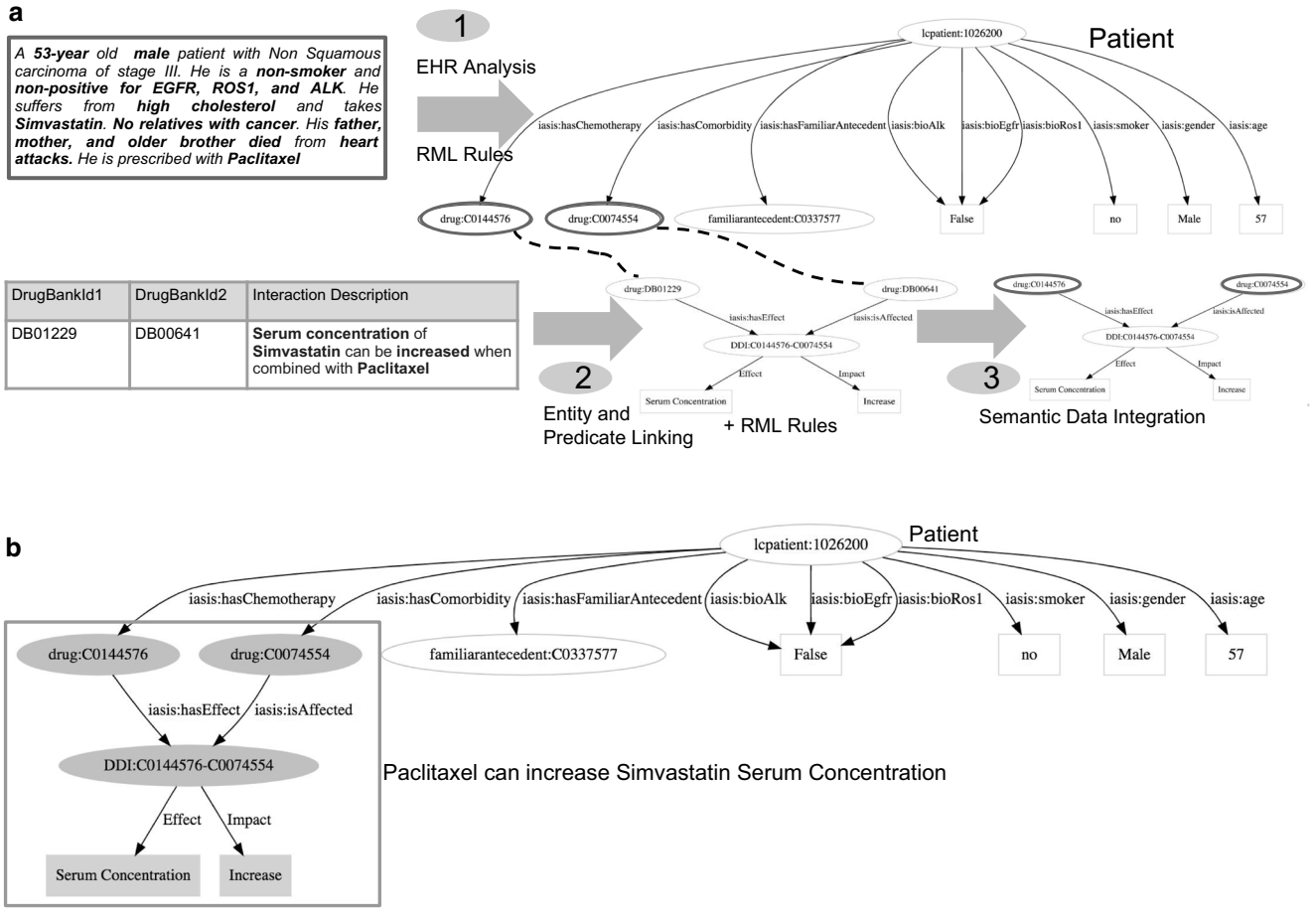


**Fig. 4** **An Example of a Pipeline for Data Integration**. **a** The pipeline receives unstructured data (step 1) and structured data sources (step 2). EHR Analysis extracts relevant facts and annotate them using UMLS terms, e.g., CO144576 and C0074554 represent Paclitaxel and Simvastatin, respectively. Entity and Predicate Linking are performed to extract the effect of drug-drug interactions on data collected from DrugBank. Mappings between identifiers of drugs in UMLS and DrugBank enable the integration of the patient with drug interactions (step 3). **b** A portion of the RDF subgraph representing a patient and the interactions of his prescribed drugs. **a** Data integration pipeline, **b** a portion of a patient in the KG

tions, lung cancer Patients, Biomarkers, Genes, Side Effects, Proteins, Enzymes, Transporters, and Annotations. The class annotation has 69,910,644 instances related to the rest of the classes in the knowledge graphs. To generate the iASiS knowledge graph, 103 RML mapping rules were defined and curated by four knowledge engineers. As a result of following the semantic data integration pipeline illustrated in Fig. 4a, the lung cancer patients were linked to the interactions between their prescribed drugs. Fig. 5 presents the density distribution of pairs of drugs that interact in the lung cancer treatment; almost 50% of the patients are taking at least one pair of drugs whose interaction has been registered by DrugBank. In average, the patients in the iASiS knowledge graph receive treatments with 1.7 reported interactions. This information is extremely valuable for the clinicians because by traversing the knowledge graph, they can easily identify the potential interactions of the drugs and prescribe more effective and less toxic treatments.

**3-Exploration and Visualization.** MULDER is a federated query engine [16] that enables the execution of queries against the federation composed by the iASiS knowledge graph, DBpedia, and Bio2RDF. MULDER receives as input, queries in SPARQL and performs the tasks of source selection, and query decomposition and optimization by exploiting the meta-data about the classes and the connections of these classes in the knowledge graphs. Moreover, MULDER relies on adaptive physical operators, e.g., symmetric join [18] and gjoin [1], and is able to produce results incrementally as soon as they are collected from the knowledge graphs. In order to illustrate the features of MULDER, we report on the results of an experiment over the complex queries of the LSLOD [25] benchmark. The state-of-the-art federated query engine ANAPSID[1] is included in the study. ANAPSID and MULDER resort to the same set of physical operators. Thus, it is expected than the differences observed between them is produced as a consequence of producing efficient plans. LSLOD [25] is

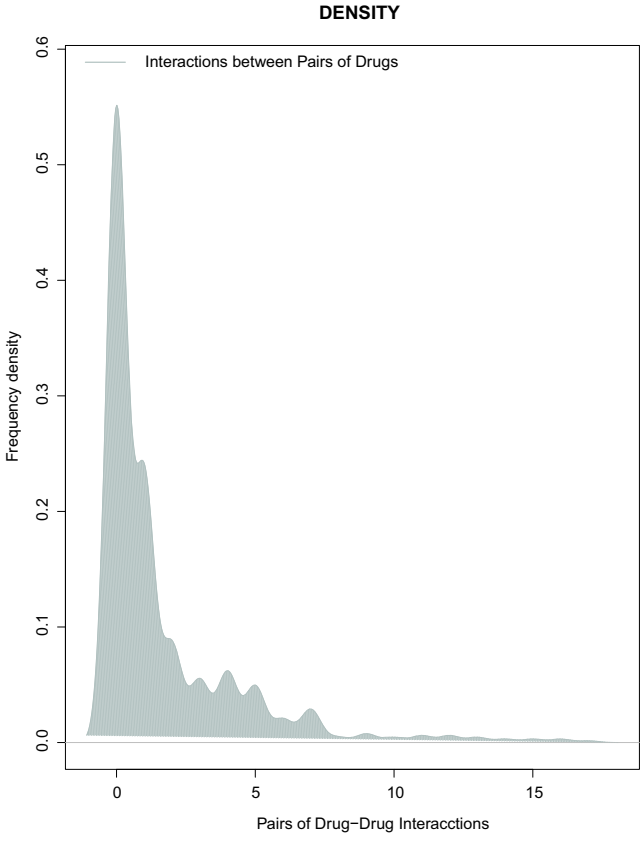


**Fig. 5 Integration of Drug Interactions**. Frequency density of pairs of drugs prescribed to patients in the knowledge graph that have known interactions (source DrugBank). As observed, there is at least one drug interaction for almost 50% of the population, and in average there are 1.7 interactions per patient

a benchmark composed of ten knowledge graphs from the life sciences domain[16]. They include: ChEBI (the Chemical Entities of Biological Interest), KEGG (Kyoto Encyclopedia of Genes and Genomes), DrugBank, TCGA-A (subset of The Cancer Genome Atlas), LinkedCT (Linked Clinical Trials), SIDER (Side Effects Resource), Affymetrix, Diseasome, DailyMed, and Medicare. The goal of the experiment is to evaluate the performance of MULDER in large data sets from the biomedical domain and in complex queries. We evaluate the efficiency in terms of the continuous generation of query answers, and use the measure *dieft@t* proposed by Acosta et al. [3]. This metric measures the continuous efficiency of an engine in the first t time units of query execution; it is computed as the AUC (area-under-the-curve) of the answer distribution until time t. Additionally, we report on multiple metrics that evaluate the overall performance and completeness, i.e., inverse of time for the first tuple ($TFFT^{-1}$), inverse of total execution time ($ET^{-1}$), number of answers (Comp), and throughput (T); all of them are "**higher is better**". Fig. 6 reports on the results of these metrics. As observed, in queries CQ2 and CQ8, ANAPSID did not produce any results before reaching the timeout (300 secs.). In the rest of the queries, both MULDER and ANAPSID are able to produce all the query answers. With the exception of CQ4 and CQ10, MULDER continuously produces results faster. Surprisingly, MULDER and ANAPSID generated the same plans for CQ4 and CQ10; however, the implementation of the physical operators impacts on a faster execution of these plans in ANAPSID. These results suggest that MULDER plans allow for a continuous performance during the answer generation process. In the context of the iASiS framework, this feature is extremely relevant because users demand to receive answers fast and continuously.

**4-Evaluation and Knowledge Discovery.** Knowledge discovery techniques are used to uncover patterns in the iASiS knowledge graph. Patterns include common characteristics of patients depending on their toxic habits, familial antecedents, or comorbities. We define a similarity measure as a function that quantifies the similarity of two patients. The patient similarity combines similarity values of the main characteristics of the two patients: age, gender, mutated genes, toxic habits, the evolution of a tumor, the mutations, and the patient performance status (ecog). Similarity values between these characteristics are computed based of different similarity measures:

(i) Lists are compared using Spearman's rho while the Jaccard similarity coefficient is utilized for sets;
(ii) similarity between drugs is computed based on the chemical structure of the drugs (SIMCOMP)[17];
(iii) side effects are compared using the Human Phenotype Ontology similarity (HPOSim)[18]; and
(iv) The UMLS similarity measure[19] is used for UMLS terms.

The combination of the similarity values is computed in terms of a triangular norm. Fig. 7a depicts the density distribution of the similarity values for pairs of lung cancer patients in the iASiS knowledge graph. We can observe that a considerably large portion of the population of patients have relatively high values of similarity, suggesting that a large number of patients have similar reactions to the prescribed treatments. Further analysis with clinical partners is required to validate the meaning of observed values of similarity. Furthermore, we apply community detection algorithms to discover patterns between patients that share similar properties in the iASiS knowledge graph. We resort to semEP (Semantics Based Edge Partitioning Problem) [45] for computing communities of patients based on the values

[16] The ten knowledge graphs have 133,873,127 RDF triples.

[17] http://www.genome.jp/tools/simcomp/.

[18] https://sourceforge.net/projects/hposim/.

[19] http://www.d.umn.edu/~tpederse/umls-similarity.html.

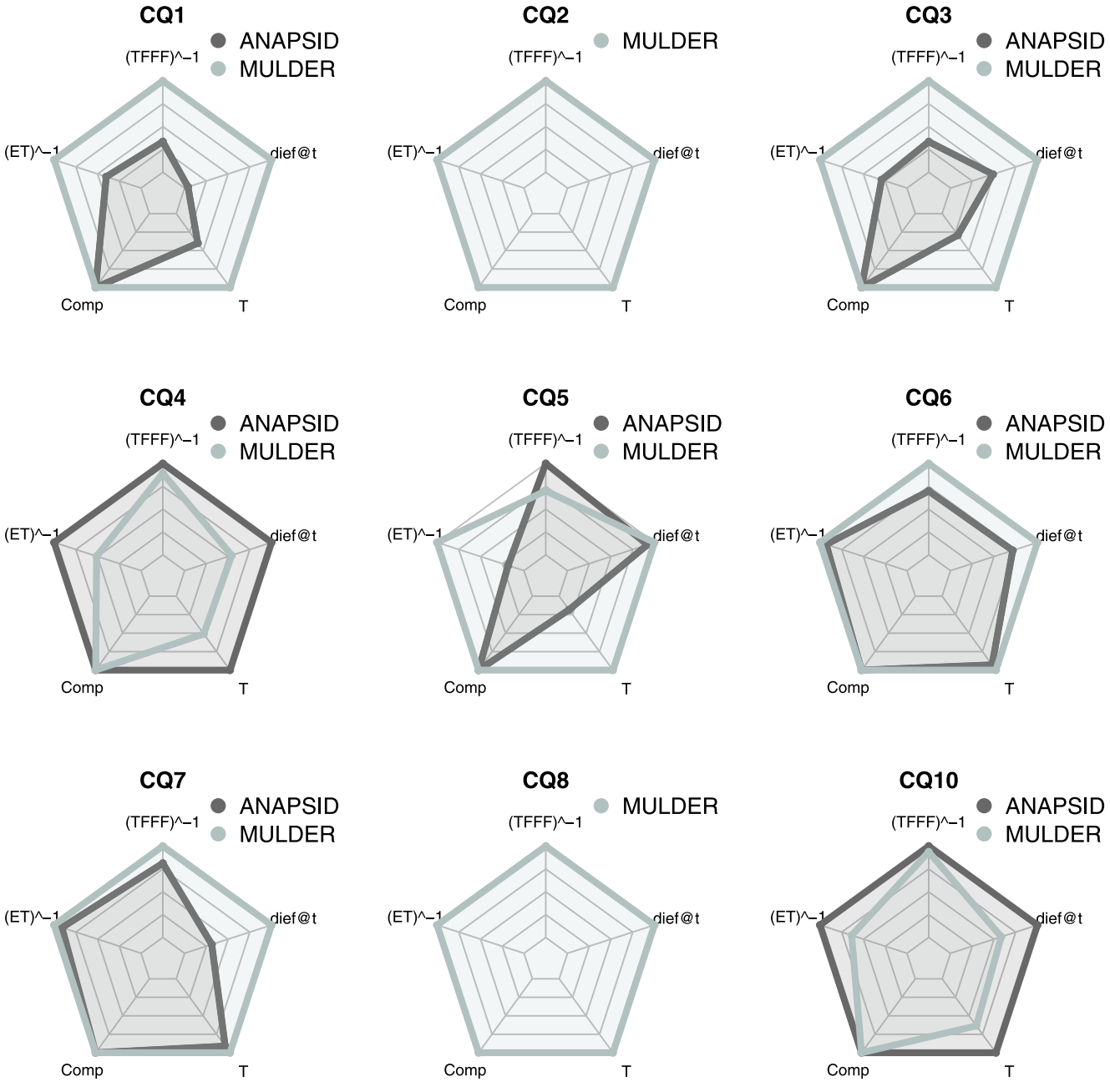


**Fig. 6** **Performance of Federated Query Engines**. ANAPSID and MULDER are compared in terms of continuous behavior. Axes correspond to: inverse of time for the first tuple ($TFFT^{-1}$), inverse of total execution time ($ET^{-1}$), number of answers produced (Comp), throughput (T), and *dief@t*. All metrics are 'higher is better'. Complex queries are from the LSLOD benchmark. ANAPSID produces empty results for CQ2 and CQ8. MULDER plans exhibit better performance than ANAPSID plans (CQ1, CQ3, CQ5, CQ6, and CQ7). Plans for CQ4 and CQ10 are the same, but ANAPSID query engine has a better continuous performance than MULDER

of similarity. It creates a minimal partitioning of the input graph, such that the density of each community is maximal. The community density represents the degree of similarity of the entities in a community. Fig. 7b reports on the results of computing semEP against the iASiS knowledge graph. Main properties of the patients involve mutations of lung cancer related genes, e.g., EGFR; demographic attributes, smoking habits, treatments, and tumor stages. The studied population is composed of 739 patients. The goal of the study is to identify the four communities of patients—out of 13 communities—with characteristics that differed from the whole population; the Kolmogorov-Smirnov test was used to rank the communities. Fig. 7b reports on four communities of patients; using a heatmap plot the percentage of patients in each community or cluster is described in terms of age, gender, EGFR mutation, and smoking habits.

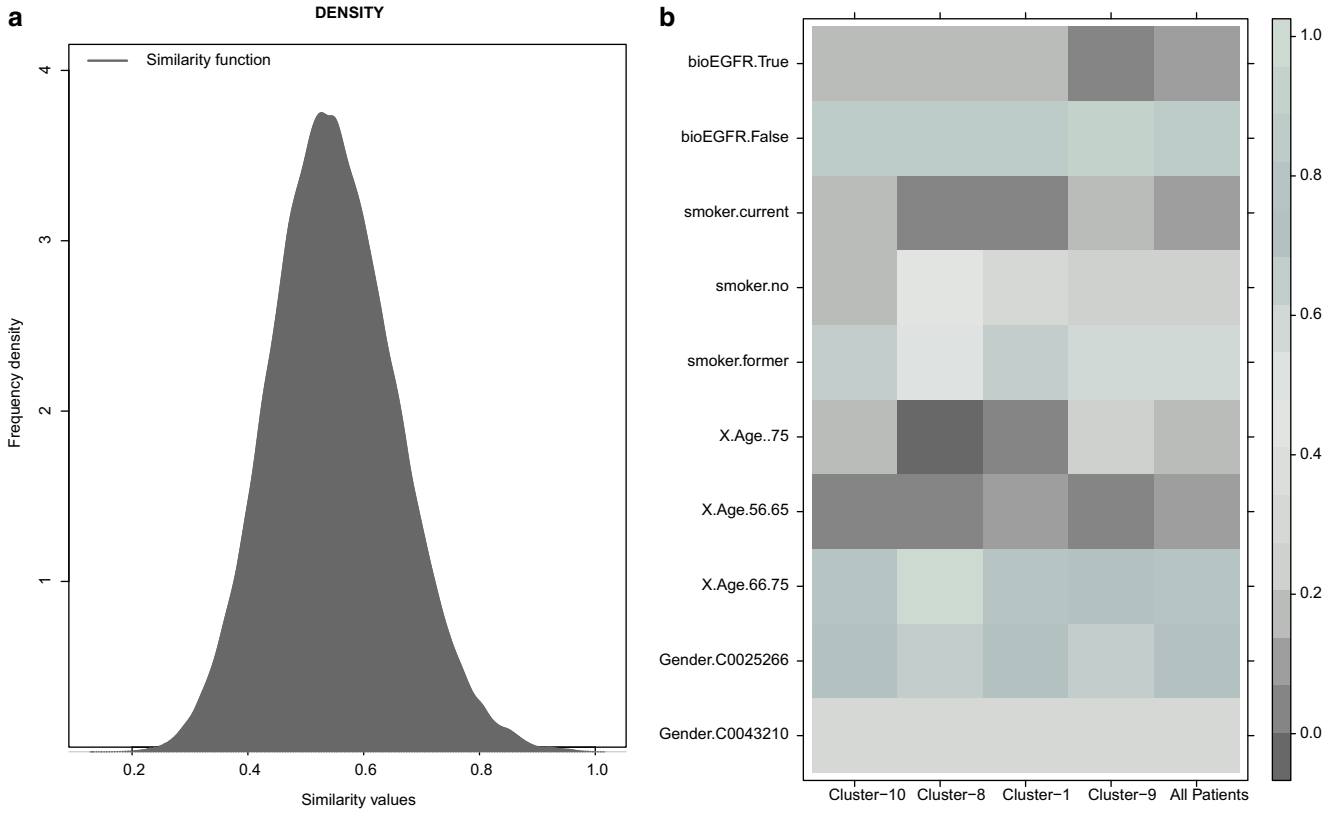


**Fig. 7** **Knowledge Analytics**. **a** A function able to quantify the similarity between two lung cancer patients is described in terms of frequency density; the function takes into account treatments, and the evolution of the tumors, mutations, and patient performance. The reported results suggest that a large number of patients react similarly to the treatments. However, more studies are required to validate this observation. **b** Communities of lung cancer patients and the summary of the observed features age, toxic habits, and EGFR mutations. Distributions of the observed features differ from the whole population, enabling the study of patients with unique characteristics. **a** Density distribution patient similarity, **b** communities of lung cancer patients

For example, patients in Cluster-1 are not current smokers and a considerably number of them are non-smokers; in addition, the biomarker EGFR is negative for many of them. The results are initial and require further study from the clinical partners of the project. However, they suggest that these techniques have the power of uncovering patterns between the observed features of patients.

## 5 Related Work

The problem of devising data integration frameworks has extensively treated in the literature [23]. The mediator and wrapper architecture proposed by Wiederhold [54] and the data integration system approach presented by Lenzerini [32], represent the basis for the state of the art [15, 24]. The community of Semantic Web have proposed various approaches that enable the integration and processing of Web data. KARMA [31] is a semi-automatic tool able to generate mapping rules between structured sources in different formats, e.g., CSV, JSON, and XML, and a unified schema. Albeit effective during the mapping definition phase, KARMA does not provide any support for the steps of data integration, curation, management, and analytics. DIG [19] and MINTE [10, 11, 18] also enable the creation of knowledge graphs, but they mainly focus on solving the problem of entity matching effectively. LDIF [6], LIMES [43], Sieve [37], Silk [29], and RapidMiner LOD Extension [49] also tackle the problem of data integration. However, they resort to similarity measures and link discovery methods to match equivalent entities from different RDF graphs. With the aim of transforming structured data in tabular or nested formats like CSV, relational, JSON, and XML, into RDF knowledge graphs, diverse mapping languages have been proposed. Exemplary mapping languages and frameworks include RDF Mapping Language (RML) [14], R2RDF [52], and R2RML [48]. Additionally, a vast amount of research has been conducted to propose effective approaches for ontology alignment [17, 39, 9], as well as to effectively perform curation of knowledge graphs [2, 35, 4]. Our knowledge-driven framework while generic, facilitates the integration of existing components; thus, it can benefit from these tools to effectively solve the problem of transforming data into actionable knowledge.

## 6 Conclusion and Future Directions

We present a knowledge-driven framework able to integrate knowledge extraction, semantic data integration, query processing, and knowledge analytics for supporting decision and policy making. We have described the application of the framework in the biomedical domain and shown the potential for uncovering patterns that can enable the explanation of treatment interactions and patient characterization. The framework is part of the iASiS platform, and clinicians are starting the process of evaluation of outcomes. Although we focus on the biomedical domain, the general knowledge-driven framework has also been applied in other domains [11], e.g., law enforcement, job market application, and smart manufacturing. Similarly, we observed that the framework is not only easy to configure, but also provides accurate results. We hope that our proposed techniques will help clinicians and data practitioners in the complex tasks of extracting valuable knowledge from heterogeneous datasets. In the future we plan to define a hybrid approach able to combine the wisdom of the domain experts and users, and the accuracy of machine learning approaches, to facilitate the evaluation of the knowledge graph and the uncovered insights.

**Acknowledgements** This work has been partially funded by the EU H2020 Project No. 727658 (IASIS).

## References

1. Acosta M, Vidal M, Lampo T, Castillo J, Ruckhaus E (2011) ANAPSID: an adaptive query processing engine for SPARQL endpoints. In: Proceedings of the 10th International Conference on The Semantic Web ISWC Bonn, 23.10.-27.10., pp 18–34 https://doi.org/10.1007/978-3-642-25073-6_2
2. Acosta M, Simperl E, Flöck F, Vidal M (2017a) Enhancing answer completeness of SPARQL queries via crowdsourcing. J Web Semant 45:41–62
3. Acosta M, Vidal M, Sure-Vetter Y (2017b) Diefficiency metrics: measuring the continuous efficiency of query processing approaches. In: The Semantic Web – ISWC 2017 – 16th International Semantic Web Conference
4. Acosta M, Zaveri A, Simperl E, Kontokostas D, Flöck F, Lehmann J (2018) Detecting linked data quality issues via crowdsourcing: a dbpedia study. Semant Web 9(3):303–335
5. Agerri R, Artola X, Beloki Z, Rigau G, Soroa A (2015) Big data for natural language processing: a streaming approach. Knowl Based Syst 79:36–42
6. Schulz A, Matteini A, Isele R, Mendes PM, Bizer C, Becker C (2012) Ldif- a framework for large-scale linked data integration. In: Proceedings of the 21st International World Wide Web Conference WWW, Developers Track Lyon, 16.04.-20.04.
7. Angles R, Arenas M, Barceló P, Hogan A, Reutter JL, Vrgoc D (2017) Foundations of modern query languages for graph databases. ACM Comput Surv 50(5):68:1–68:40
8. Ceri S, Gottlob G, Tanca L (1989) What you always wanted to know about datalog (and never dared to ask). IEEE Trans Knowl Data Eng 1(1):146–166
9. Cheatham M, Cruz IF, Euzenat J, Pesquita C (2017) Special issue on ontology and linked data matching. Semant Web 8(2):183–184
10. Collarana D, Galkin M, Ribón IT, Vidal M, Lange C, Auer S (2017) MINTE: semantically integrating RDF graphs. In: Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, WIMS 2017 Amantea, 19.06.-22.06.. https://doi.org/10.1145/3102254.3102280
11. Collarana D, Galkin M, Lange C, Scerri S, Auer S, Vidal M (2018) Synthesizing knowledge graphs from web sources with the MINTE++ framework. In: The Semantic Web – ISWC 2018 – 17th International Semantic Web Conference
12. Cruz AL, Baranya A, Vidal M (2012) Medical image rendering and description driven by semantic annotations. In: Resource Discovery – 5th International Workshop, RED 2012, Co-located with the 9th Extended Semantic Web Conference, ESWC 2012 Heraklion, 27.05.2012, pp 123–149 (Revised Selected Papers)
13. Daiber J, Jakob M, Hokamp C, Mendes PN (2013) Improving efficiency and accuracy in multilingual entity extraction. In: I-SEMANTICS 2013 – 9th International Conference on Semantic Systems, ISEM '13 Graz, 04.09.-06.09., pp 121–124
14. Dimou A, Sande MV, Colpaert P, Verborgh R, Mannens E, de Walle RV (2014) RML: a generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014)
15. Doan AH, Halevy AY, Ives ZG (2012) Principles of Data Integration. Morgan Kaufmann, ISBN 978-0-12-416044-6, pp I–XVIII, 1–497
16. Endris KM, Galkin M, Lytra I, Mami MN, Vidal M, Auer S (2018) Querying interlinked data by bridging RDF molecule templates. T Large Scale Data Knowl Cent Syst 39:1–42
17. Euzenat J, Shvaiko P (2013) Ontology matching, 2nd edn. Springer, Berlin Heidelberg
18. Galkin M, Collarana D, Ribón IT, Vidal M, Auer S (2017) Sjoin: A semantic join operator to integrate heterogeneous RDF graphs. In: Database and Expert Systems Applications – 28th International Conference, DEXA 2017 Lyon, 28.08.-31.08., pp 206–221 (Proceedings, Part I)
19. Gawriljuk G, Harth A, Knoblock CA, Szekely PA (2016) A scalable approach to incrementally building knowledge graphs. In: Research and Advanced Technology for Digital Libraries – 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016 Hannover, 05.09.-09.09., pp 188–199 (Proceedings)
20. Getoor L (2013) Probabilistic soft logic: a scalable approach for markov random fields over continuous-valued variables – (abstract of keynote talk). In: Theory, Practice, and Applications of Rules on the Web – 7th International Symposium, RuleML 2013 Seattle, 11.07.-13.07., p 1 (Proceedings)
21. Golshan B, Halevy AY, Mihaila GA, Tan W (2017) Data integration: after the teenage years. In: Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017 Chicago, 14.05.-19.05., pp 101–106
22. Halevy AY (2017) Technical perspective: building knowledge bases from messy data. Commun ACM 60(5):92
23. Halevy AY (2018) Information integration. In: Encyclopedia of Database Systems, 2nd edn.
24. Halevy AY, Rajaraman A, Ordille JJ (2006) Data integration: the teenage years. In: Proceedings of the 32nd International Conference on Very Large Data Bases Seoul, 12.09.-15.09., pp 9–16
25. Hasnain A, Mehmood Q, Sana E, Zainab S, Saleem M, Warren C, Zehra D, Decker S, Rebholz-Schuhmann D (2017) Biofed: federated query processing over life sciences linked open data. J Biomed Semantics 8(1):13
26. Hassanzadeh O, Chiang F, Miller RJ, Lee HC (2009) Framework for evaluating clustering algorithms in duplicate detection. Proceedings VLDB Endowment 2(1):1282–1293

27. Henning CA, Ewerth R (2018) Estimating the information gap between textual and visual representations. Int J Multimed Inf Retr 7(1):43–56
28. Hu W, Qiu H, Huang J, Dumontier M (2017) Biosearch: a semantic search engine for bio2rdf. Database. https://doi.org/10.1093/database/bax059
29. Isele R, Bizer C (2013) Active learning of expressive linkage rules using genetic programming. J Web Semant 23:2–15. https://doi.org/10.1016/j.websem.2013.06.001
30. Klimchuk OI, Konovalov KA, Perekhvatov VV, Skulachev KV, Dibrova DV, Mulkidjanian AY (2017) Cognat: a web server for comparative analysis of genomic neighborhoods. Biol Direct. https://doi.org/10.1186/s13062-017-0196-z
31. Knoblock CA, Szekely PA (2015) Exploiting semantics for big data integration. AI Mag 36(1):25–38
32. Lenzerini M (2002) Data Integration: a theoretical perspective. In: Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems Madison, 03.06.-05.06., pp 233–246
33. Libkin L, Reutter JL, Soto A, Vrgoc D (2018) TriAL: A navigational algebra for RDF triplestores. Acm Trans Database Syst 43(1):5:1–5:46
34. Livi CM, Klus P, Delli Ponti R, Tartaglia GG (2016) catrapid signature: identification of ribonucleoproteins and rna-binding regions. Bioinformatics 32(5):773–775. https://doi.org/10.1093/bioinformatics/btv629
35. Loster M, Naumann F, Ehmueller J, Feldmann B (2018) Curex: a system for extracting, curating, and exploring domain-specific knowledge graphs from text. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018 Torino, 22.10.-26.10.
36. Menasalvas E, González AR, Costumero R, Ambit H, Gonzalo C (2016) Clinical narrative analytics challenges. In: Rough Sets – International Joint Conference, IJCRS 2016 Santiago de Chile, 07.10.-11.10., pp 23–32 (Proceedings)
37. Mendes PN, Mühleisen H, Bizer C (2012) Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops Berlin, 30.03., pp 116–123
38. Ross MK, Wei W, Ohno-Machado L (2014) Big data and the electronic health record. IMIA yearbook of medical Informatics, vol 1
39. Mohammadi M, Atashin AA, Hofman W, Tan Y (2018) Comparison of ontology alignment systems across single matching task via the mcNemar's test. TKDD 12(4):51:1–51:18
40. Munevar S (2017) Unlocking big data for better health. Nat Biotechnol 35(7):684–686. https://doi.org/10.1038/nbt.3918
41. Navigli R (2018) Natural language understanding: instructions for (present and future) use. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018 Stockholm, 13.07.-19.07., pp 5697–5702
42. Nentidis A, Bougiatiotis K, Krithara A, Paliouras G (2018) Semantic integration of disease-specific knowledge. In: Poster in European Conference on Computational Biology (ECCB18)
43. Ngomo ACN, Auer S (2011) Limes-a time-efficient approach for large-scale link discovery on the web of data. In: IJCAI, pp 2312–2317
44. Ortiz CA, Gonzalo-Martín C, Garcia-Pedrero A, Ruiz EM (2018) Supervoxels-based histon as a new alzheimer's disease imaging biomarker. Sensors 18(6):1752
45. Palma G, Vidal M, Raschid L (2014) Drug-target interaction prediction using semantic similarity and edge partitioning. In: ISWC
46. Papachristou N, Puschmann D, Barnaghi P, Cooper B, Hu X, Maguire R, Apostolidis K, Conley YP, Hammer M, Katsaragakis S, Kober KM, Levine JD, McCann L, Patiraki E, Furlong EP, Fox PA, Paul SM, Ream E, Wright F, Miaskowski C (2018) Learning from data to predict future symptoms of oncology patients. PLoS ONE. https://doi.org/10.1371/journal.pone.0208808
47. Perez W, Tello A, Saquicela V, Vidal M, Cruz AL (2015) An automatic method for the enrichment of DICOM metadata using biomedical ontologies. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015 Milan, 25.08.-29.08., pp 2551–2554
48. Priyatna F, Corcho Ó, Sequeda JF (2014) Formalisation and experiences of R2RML-based SPARQL to SQL query translation using morph. In: 23rd International World Wide Web Conference, WWW '14 Seoul, 07.04.–11.04., pp 479–490
49. Ristoski P, Bizer C, Paulheim H (2015) Mining the web of linked data with rapidminer. Web Semant 35:142–151
50. Ruiz EM, Tuñas JM, Bermejo G, Gonzalo-Martín C, González AR, Zanin M, de Pedro CG, Mendez M, Zaretskaia O, Rey J, Parejo C, Bermudez JLC, Provencio M (2018) Profiling lung cancer patients using electronic health records. J Med Syst 42(7):126:1–126:10
51. Sakor A, Mulang' IO, Singh K, Shekarpour S, Vidal ME, Lehmann J, Auer S (2019) Old is gold: linguistic driven approach for entity and relation linking of short text. In: Proceedings of the NAACL HLT
52. Sequeda JF, Arenas M, Miranker DP (2014) OBDA: query rewriting or materialization? in practice, both! In: The Semantic Web – ISWC 2014 – 13th International Semantic Web Conference Riva del Garda, 19.10.-23.10., pp 535–551 (Proceedings, Part I)
53. Tukiainen T (2017) Landscape of x chromosome inactivation across human tissues. Nature. https://doi.org/10.1038/nature24265
54. Wiederhold G (1992) Mediators in the architecture of future information systems. IEEE Comput 25(3):38–49
55. Zadorozhny V, Raschid L, Vidal M, Urhan T, Bright L (2002) Efficient evaluation of queries in a mediator for websources. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data Madison, 03.06.-06.06., pp 85–96
56. Zhong RY, Newman ST, Huang GQ, Lan S (2016) Big data for supply chain management in the service and manufacturing sectors: challenges, opportunities, and future perspectives. Comput Ind Eng 101:572–591