Open Research Online

The Open University's repository of research publications
and other research outputs

Enriching Data Lakes with Knowledge Graphs

Conference or Workshop Item

How to cite:

Chessa, Alessandro; Fenu, Gianni; Motta, Enrico; Reforgiato Recupero, Diego; Osborne, Francesco; Salatino, Angelo and Secchi, Luca Enriching Data Lakes with Knowledge Graphs. In: Knowledge Graph Generation from Text.

For guidance on citations see FAQs.

© [not recorded]

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Accepted Manuscript

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Enriching Data Lakes with Knowledge Graphs

Alessandro Chessa[1,2], Gianni Fenu[3][0000−0003−4668−2476], Enrico Motta[4][0000−0003−0015−1952], Francesco Osborne[4,5][0000−0001−6557−3131], Diego Reforgiato Recupero[3][0000−0001−8646−6183], Angelo Salatino[4][0000−0002−4763−3943], and Luca Secchi[1,3]

[1] Linkalab s.r.l., Cagliari, Italy {alessandro.chessa, luca.secchi}@linkalab.it
[2] Luiss Data Lab, Rome, Italy
[3] University of Cagliari, Cagliari, Italy {fenu,diego.reforgiato}@unica.it
[4] Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom {enrico.motta,francesco.osborne,angelo.salatino}@open.ac.uk
[5] University of Milano Bicocca, Milan, Italy

Abstract. Data lakes are repositories of data stored in natural/raw format. A data lake may include structured data from relational databases, semi-structured data (i.e., JSON, CSV), unstructured data (i.e., text data), or binary data (i.e., images, audio, video). It is usually built on top of cost-efficient infrastructures such as Hadoop, Amazon S3, MongoDB, ElasticSearch, etc. Several organisations rely on big data lakes for crucial tasks such as reporting, visualisation, advanced analytics, machine learning, and business intelligence. A major limitation of this solution is that without descriptive metadata and a mechanism to maintain it, such data tend to be noisy, making their management and analysis complex and time-consuming. Therefore, there is the need to add a semantic layer based on a formal ontology to describe the data and efficient mechanism to represent them as a knowledge graph. In this paper, we present a methodology to add a semantic layer to a data lake and thus obtain a knowledge graph that can support structured queries and advanced data exploration. We describe a practical implementation of a methodology applied to a data lake consisting of text data describing the online marketplace for lodging and tourism activities. We report statistics about the data lake and the resulting knowledge graph.

Keywords: Semantic Data Lake · Knowledge Graphs · Information Extraction · Data Mining

1 Introduction

The term "data lake" was introduced by James Dixon, Chief Technology Officer of Pentaho, in a blog post in 2010[6]. Data lakes are data repositories for storing large and heterogeneous sets of raw data. They have quickly become a common data management solution for organizations that desire to own a holistic and large repository for their data. Data lakes allow users to access and explore data

[6] https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

without the need to move them into another system. Insights and reporting performed from a data lake typically occur on an ad-hoc basis. However, users might apply a schema and a certain degree of automation to the data to make it possible to duplicate a report when needed.

Data in a data lake are stored in their raw format and are not transformed until they are needed for analysis. Also in such a case, a schema is somehow applied so that they can be analyzed. This way of working is called "schema on read" [24, 17], because data are kept raw until they are ready to be used. Although it is always possible to use a schema-on-read approach, it is not optimal for performances and optimising costs, thus sometimes data is transformed and stored using specific file formats (e.g., Parquet, AVRO, ORC) that can handle also schema information. Data lakes require governance to establish continual maintenance and to keep the data usable and accessible. Otherwise, the risk is to end up with data which become inaccessible, unwieldy, expensive, and useless; culminating to what is often referred as "data swamps"[7]. In order to address this limitation, it is useful to rely on a semantic layer: a representation of data based on semantic technologies and a formal ontology that can offer a unified, consolidated view of data across the organisation.

Several attempts have been done to provide a semantic layer to data lakes and each of them has targeted a particular domain of application [6, 11, 4, 12, 14, 13, 19, 22]. However, to the best of our knowledge, no one has ever applied this solution in the domain of tourism. In this paper, we propose a practical implementation for the creation of a semantic layer to generate a knowledge graph from a data lake consisting of text data. We applied this solution in the tourism domain, developing a knowledge graph of accommodation facilities in London, leveraging the *Data Lake Turismo* platform. Our solution takes advantage of entity linking approaches for extracting and interlinking several entities (e.g., places, food, amenities) from reviews and other textual fields, allowing a much more comprehensive representation of accommodations and touristic locations. This *Data Lake Turismo* was developed by Linkalab s.r.l.[8], capitalising on a previous research project promoted by the Digital Innovation Hub of Sardinia[9] and Fondazione di Sardegna[10].

The remainder of this paper is organised as it follows. Section 2 focuses on the previous works on semantic layers for data lakes. Section 3 describes our methodology and presents the implementation in the tourism domain. We provide also statistics and information about the resulting knowledge graph. Section 4 ends the paper with conclusions and future works.

[7] https://developer.ibm.com/articles/ba-data-becomes-knowledge-2/
[8] Linkalab s.r.l. is a Italian small enterprise specialised in data science and data engineering. Home page https://www.linkalab.it/
[9] https://www.dihsardegna.eu/
[10] https://www.fondazionedisardegna.it/

2 Related Work

A knowledge graph [10, 18, 3, 2, 21, 20, 1] is a knowledge base that uses a graph-structured data model to integrate data. It represents a network of real-world entities, i.e., objects, events, situations, or concepts, and illustrates the relationship between them.

Dibowski et al. [12] discussed how to address data findability, accessibility, interoperability, and re-use for data stored in a data lake. They showed the benefits provided to a data lake through the support of ontologies and knowledge graphs which provide cataloguing of data, tracking provenance, access control, and semantic search. In particular, they built the DCPAC ontology (Data Catalog, Provenance, and Access Control) related to the management of data produced by vehicles. Similarly, Diamantini et al. [11] presented a semantic model for the correct data fruition stored into a data lake. They mapped the indicators of interest, the dimensions of analysis and formulas into a knowledge graph to support the correct identification of data. Pomp et al. [22] had similar problems related to the collection, finding, understanding and accessing of large data sources with the goal of ensuring their real-time availability. To reduce the time from the collection to the analysis of data, they centralised the data in a data lake. Instead of populating the data lake of unstructured data, they proposed a semantic data platform called ESKAPE for the semantic annotation of the ingested data. Furthermore, a knowledge graph has been defined to act as an index that evolves over time according to the data that are included. In this way, users can easily identify and analyse the data coming from the different places. Bagozi et al. [6] proposed a semantics-based approach for the personalised exploration of data lakes within the domain of smart cities. First, they provided the data lake with a semantic model using domain ontologies. Then, another ontology was adopted to describe indicators and analysis dimensions. Finally, personalised exploration graphs were generated for different types of users. Another work worth mentioning is by Ansari et al. [4], who proposed a semantic profiling tool for metadata extension in data lake systems. Its aim was to understand the meaning of data. Their tool recognised the meaning of data at schema and instance level using domain vocabularies and ontologies. Finally, Mami [19] proposed a physical and a logical data integration whose goal was to query large and heterogeneous data sources. For the physical data integration they defined an ontology to transform the data into RDF.

Differently from the approaches above, we propose a methodology to extend a data lake containing data extracted from touristic platforms with a semantic layer and produce a knowledge graph. To this end, we engineered an ontology in the touristic domain integrating already existing ontologies and extending them with our classes. However, the focus of this manuscript is not on the ontology but on the extracted knowledge graph and the steps we performed to transform the data from the data lake to the knowledge graph.

3 The Proposed Methodology

In this section, we describe our methodology for enriching a data lake by creating a domain ontology and generating a knowledge graph that will extend the data lake with a sophisticated representation of knowledge. This approach is articulated in five steps: i) analysis of the data sources; ii) definition of the use cases; iii) creation of the ontology; iv) data transformation; and v) generation of the knowledge graph. In the following we will briefly describe each phase.

3.1 Analysis of the data sources

The data lake we have used comes as a result of the *Data Lake Turismo*[11] project whose aim was to create a digital platform of tourism data. The data lake was developed by Linkalab[12] through Amazon Web Services (AWS) cloud computing technologies including S3 where the data was stored. The data lake collected data from various sources using an Extraction, Load and Transformation (ELT) approach. A crawling system was developed to identify and extract data related to the London Region[13] area from various sources including Booking.com[14] and AirBnB[15].

The data lake is organised in three tiers: i) *intake tier*, where the raw data is collected, ii) *curated tier*, where each transformed/cleaned version of the data is stored, and iii) *consumption tier*, where the data is exposed to business analysts in many formats such as reports, dashboards, APIs. In our specific scenario, the intake tier contains HTML files extracted from Booking.com, and JSON files extracted using APIs exposed by AirBnB systems; the curated tier contains JSON data extracted from the Booking.com HTML files; the consumption tier contains the knowledge graph as a set of RDF triples.

The data lake is built on AWS serverless technologies: Amazon S3[16] object storage is used to store the files, AWS Lambda[17] and AWS Fargate[18] are used to execute the crawling and the data processing, Amazon Athena[19] is used to query the data stored in JSON files using SQL language while all technical metadata is managed using AWS Glue catalog[20].

The data lake describes three kinds of entities:

[11] Turismo means tourism in Italian.
[12] Linkalab - https://www.linkalab.it/
[13] The London Region area is an administrative area including the 32 London boroughs and the City of London.
[14] https://www.booking.com/
[15] https://www.airbnb.com/
[16] See https://aws.amazon.com/s3/
[17] See https://aws.amazon.com/lambda/
[18] See https://aws.amazon.com/fargate/
[19] Athena is a query engine based on PrestoDB. See https://aws.amazon.com/athena/ and https://prestodb.io/
[20] See https://aws.amazon.com/glue/

Table 1. Entities stored in the data lake referring to the London Region (UK).

Source	Zone	Entity type	Total number
Booking.com	London	Lodging facility	2,092
Booking.com	London	Accommodation offer	22,154
Booking.com	London	Review	443,675
AirBnB	London	Lodging facility	5,975
AirBnB	London	Accommodation offer	5,975
AirBnB	London	Review	142,500

- **lodging facilities** i.e., any hotel, holiday house or other quarters that provide temporary sleeping facilities open to the public[21], which are described by specific properties like name, address, geolocation, average user rating, textual description, pictures, related amenities;
- **accommodation offers** i.e., a specific place that can accommodate persons (e.g. a hotel room, a camping pitch or an entire apartment) that is part of a lodging facility and is offered for lease under specific conditions; these offers are characterised by specific properties like number and type of beds, max and min occupancy, related amenities, price;
- **user reviews** about the lodging facility that are characterised by a rating value and a text.

Table 1 reports an overview of the number of business entities stored for both sources (Booking.com and AirBnb). For AirBnb we have the same amount of lodging facilities and accommodation offers, because AirBnB associates each offer to a unique lodging facility. Conversely, in Booking.com a lodging facility (e.g., hotel) can offer multiple accommodations (e.g., rooms).

Table 2. Storage space in the data lake.

Source	Zone	Total size html	Total size json
Booking.com	London	13.6 GB	325.4 MB
AirBnB	London	-	31.6 MB

Table 2 summarize of the storage space used in the data lake. The main difference between Booking.com and AirBnB is that the first is crawled exporting HTML pages that are then used to extract the data whereas the latter is accessed using APIs to retrieve the data itself already in JSON format.

3.2 Definition of the use cases

The purpose of the creation of the *Data Lake Turismo* project was to analyse the supply and demand side of tourist destinations. During the project development

[21] Source: Law Insider, see https://www.lawinsider.com/dictionary/lodging-facilities

the following use cases have been identified in collaboration with the analysts of Linkalab:

1. Identify the topics of interest in the tourists' reviews;
2. Identify the topics of interest in the text presentations of lodging businesses offers;
3. Detect the sentiment [23] of tourists toward a certain lodging business or destinations;
4. Classify tourist destinations according to what they offer and according to the tourist opinions.

To better support these use cases the data lake has been extended with a semantic layer supported by an ontology. The resulting knowledge graph includes both data and metadata, hence enhancing the support for developing dedicated services.

3.3 Creation of the ontology

A crucial step is the creation of a domain ontology that could support the use cases. For this purpose is possible to rely on standard ontology engineering frameworks and evaluation methodologies [7].

In our implementation, the ontology has to satisfy both functional and non functional requirements. As far as functional requirements are concerned, the ontology has to include classes for lodging businesses (e.g., hotels, hostels, apartments), accommodations offered by them (e.g., rooms, suite), amenities for tourists, tourist attractions and points of interest, inter-relations among entities (e.g., geographic relations, composition/inclusion), tourist reviews, tourist destinations and taxonomies to support all of them. As far as the non functional requirements are concerned the ontology must be defined in OWL, and be based on `Schema.org`[22] and GoodRelations[23].

To drive the creation of the ontology, we designed a set of competency questions and identified a set of existing ontologies that have been used as support. The entire ontology creation is not discussed in this manuscript because out of the scope of the paper which focuses on the methodology for the creation of a knowledge graph to support a data lake.

3.4 Data transformation

The data transformation depends on the source data structures and on the desired output. The steps needed to transform the data are: i) extraction of relevant structured data and texts from the original sources; ii) data cleaning; iii) ontology mappings, to represent the entities in the structured data according to the ontology; iv) language detection, to identify the source of the language; v) identification and extraction of entities within the text.

[22] https://schema.org/
[23] http://www.heppnetz.de/projects/goodrelations/

The last step is very crucial to obtain a good representation of the data, since many important information are only expressed in natural language, especially in the text regarding the description of the lodging facilities and reviews. To this purpose, we used DBpedia Spotlight entity linking approach for extracting common entities such as activities, events, places, and food.

We then integrated this information in the knowledge graph by linking DBpedia entities with the relevant lodging facilities. This allows our system to support advanced queries such as retrieving all the accommodations that are close to touristic attractions, those that offer a specific amenity or propose a special kind of food, but also looking for what places or events users cite most frequently in their reviews.

3.5 Generation of the knowledge graph

The last step takes in input the refined data and the ontology and produced the knowledge graph. To this purpose it is possible to rely on several languages and tools for the the automatic generation of triples [9, 5]. In our implementation, we adopted the RDF Mapping Language (RML) [14], which is one of the most well-known solutions in this space, to build specific data pipelines for the creation of RDF triples. The RML language specifies how linked data are produced from the corresponding data sources. To perform an RML transformation[24] we need three things: i) an RML processor; ii) an input data source; iii) a mapping from any (structured) data in the input data source to RDF.

Triples are generated for each of the triples maps of the RML mapping. In our prototype, we used RMLMapper [13][25] for such a purpose. The triples representing the knowledge graph are generated as N-Quads files[26] which are stored in the consumption tier of the data lake.

We ingest new data from the original sources into the data lake every two months. We then recreate the knowledge graph from scratch by repeating all the data transformations steps described in Section 3.4.

Table 3 reports some metrics about the last version of the knowledge graph: i) **total statements** refers to the overall number of triples stored in the triplestore (both explicit and inferred), ii) **explicit statements** refers to the number of raw triples created in the triplestore, iii) **inferred statements** refers to the number of triples inferred by the reasoner from the explicit statements, iv) **expansion ratio** represents the percentage of triples added using the inference. The other metrics are self explanatory.

4 Conclusions

In this paper we have presented a general methodology for extending a data lake with a knowledge graph. In particular, we have focused our analysis to the

[24] https://rml.io/specs/rml/
[25] https://github.com/RMLio/rmlmapper-java
[26] See https://www.w3.org/TR/n-quads/

Table 3. Knowledge graph metrics

Metric	Value
Total statements	10,299,471
Explicit statements	5,148,987
Inferred statements	5,150,484
Expansion ratio	2
Number of distinct relations	50
Number of DBPedia entities linked	91,284
Number of unique DBpedia enities linked	2,644
Number of AirBnB reviews entities	142,500
Number of Booking.com reviews entities	435,276
Total number of reviews entities	577,776
Total time for triple generation	∼19 minutes

tourism domain by considering a data lake containing structured and unstructured data crawled from Booking.com and AirBnB. The knowledge graph thus obtained has been stored into a triplestore which can be accessed online.

We can conclude that the semantic layer provided by the knowledge graph brought many advantages to Linkalab's data lake platform: i) it treats data and metadata in a unified way, ii) it has a flexible schema that can support the data variety and evolution, iii) it supports algorithms and applications development and data science activities based on the data lake; iv) it embeds information in its graph structure that can be leveraged by graph analytics [15, 8] and representation learning [16] algorithms; v) it incorporates knowledge extracted from texts along with structured and semi-structured data typically found in the data lake; vi) it can be used to expand the data lake information context through connections with open knowledge graphs like DBpedia.

In future work, we aim to expand the pipeline for producing the knowledge graph by developing new solutions for entity extraction and to further improve the ontology. We also plan to develop a tool that will take advantage of the knowledge graph for analysing and comparing accommodations and generating explainable recommendations.

References

1. Alam, M., Fensel, A., Gil, J.M., Moser, B., Recupero, D.R., Sack, H.: Special issue on machine learning and knowledge graphs. Future Gener. Comput. Syst. **129**, 50–53 (2022). https://doi.org/10.1016/j.future.2021.11.022, https://doi.org/10.1016/j.future.2021.11.022
2. Alam, M., Gangemi, A., Presutti, V., Recupero, D.R.: Semantic role labeling for knowledge graph extraction from text. Prog. Artif. Intell. **10**(3), 309–320 (2021). https://doi.org/10.1007/s13748-021-00241-7, https://doi.org/10.1007/s13748-021-00241-7

3. Angioni, S., Salatino, A., Osborne, F., Recupero, D.R., Motta, E.: Aida: A knowledge graph about research dynamics in academia and industry. Quantitative Science Studies pp. 1–43 (2021)
4. Ansari, J.W., Karim, N., Ansari, W., Beyan, O.D., Cochez, M.: Semantic profiling in data lake (2018)
5. Arenas-Guerrero, J., Scrocca, M., Iglesias-Molina, A., Toledo, J., Gilo, L.P., Dona, D., Corcho, O., Chaves-Fraga, D.: Knowledge graph construction with r2rml and rml: an etl system-based overview (2021)
6. Bagozi, A., Bianchini, D., De Antonellis, V., Garda, M., Melchiori, M.: Personalised exploration graphs onÂ semantic data lakes. In: Panetto, H., Debruyne, C., Hepp, M., Lewis, D., Ardagna, C.A., Meersman, R. (eds.) On the Move to Meaningful Internet Systems: OTM 2019 Conferences. pp. 22–39. Springer International Publishing, Cham (2019)
7. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Pattern-based design applied to cultural heritage knowledge graphs **12**, 313–357 (2021). https://doi.org/10.3233/SW-200422
8. Cuzzocrea, A., Song, I.Y.: Big graph analytics: The state of the art and future research agenda. DOLAP 2014 - Proceedings of the ACM 17th International Workshop on Data Warehousing and OLAP, co-located with CIKM 2014 pp. 99–101 (2014). https://doi.org/10.1145/2666158.2668454
9. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E.: Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain. Future Gener. Comput. Syst. **116**, 253–264 (2021). https://doi.org/10.1016/j.future.2020.10.026, https://doi.org/10.1016/j.future.2020.10.026
10. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., Sack, H.: Ai-kg: an automatically generated knowledge graph of artificial intelligence. In: International Semantic Web Conference. pp. 127–143. Springer (2020)
11. Diamantini, C., Potena, D., Storti, E.: A semantic data lake model for analytic query-driven discovery. In: The 23rd International Conference on Information Integration and Web Intelligence. p. 183–186. iiWAS2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3487664.3487783, https://doi.org/10.1145/3487664.3487783
12. Dibowski, H., Schmid, S., Svetashova, Y., Henson, C., Tran, T.: Using semantic technologies to manage a data lake: Data catalog, provenance and access control (11 2020)
13. Dimou, A., De Nies, T., Verborgh, R., Mannens, E., de Walle, R.: Automated Metadata Generation for Linked Data Generation and Publishing Workflows. Proceedings of the 9th Workshop on Linked Data on the Web **1593** (2016)
14. Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., Van De Walle, R.: RML: A generic language for integrated RDF mappings of heterogeneous data. In: CEUR Workshop Proceedings. vol. 1184 (2014)
15. Iosup, A., Hegeman, T., Ngai, W.L., Heldens, S., Pérez, A.P., Manhardt, T., Chafi, H., Capotă, M., Sundaram, N., Anderson, M., Tănase, I.G., Xia, Y., Nai, L., Boncz, P.: LDBC graphalytics: A benchmark for large scale graph analysis on parallel and distributed platforms. Proceedings of the VLDB Endowment **9**(13), 1317–1328 (2015). https://doi.org/10.14778/3007263.3007270
16. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Transactions on Neural Networks and Learning Systems pp. 1–26 (2021). https://doi.org/10.1109/TNNLS.2021.3070843

17. Mathis, C.: Data lakes. Datenbank-Spektrum **17**(3), 289–293 (2017)
18. Meloni, A., Angioni, S., Salatino, A.A., Osborne, F., Recupero, D.R., Motta, E.: Aida-bot: A conversational agent to explore scholarly knowledge graphs. In: Seneviratne, O., Pesquita, C., Sequeda, J., Etcheverry, L. (eds.) Proceedings of the ISWC 2021 Posters, Demos and Industry Tracks: From Novel Ideas to Industrial Practice co-located with 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24-28, 2021. CEUR Workshop Proceedings, vol. 2980. CEUR-WS.org (2021), http://ceur-ws.org/Vol-2980/paper310.pdf
19. Mohamed Nadjib Mami: Strategies for a Semantified Uniform Access to Large and Heterogeneous Data Sources. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn (Feb 2021), https://hdl.handle.net/20.500.11811/8925
20. Nayyeri, M., Cil, G.M., Vahdati, S., Osborne, F., Kravchenko, A., Angioni, S., Salatino, A.A., Recupero, D.R., Motta, E., Lehmann, J.: Link prediction of weighted triples for knowledge graph completion within the scholarly domain. IEEE Access **9**, 116002–116014 (2021). https://doi.org/10.1109/ACCESS.2021.3105183, https://doi.org/10.1109/ACCESS.2021.3105183
21. Nayyeri, M., Cil, G.M., Vahdati, S., Osborne, F., Rahman, M., Angioni, S., Salatino, A.A., Recupero, D.R., Vassilyeva, N., Motta, E., Lehmann, J.: Trans4e: Link prediction on scholarly knowledge graphs. Neurocomputing **461**, 530–542 (2021). https://doi.org/10.1016/j.neucom.2021.02.100, https://doi.org/10.1016/j.neucom.2021.02.100
22. Pomp, A., Paulus, A., Kirmse, A., Kraus, V., Meisen, T.: Applying semantics to reduce the time to analytics within complex heterogeneous infrastructures. Technologies **6**(3) (2018). https://doi.org/10.3390/technologies6030086, https://www.mdpi.com/2227-7080/6/3/86
23. Reforgiato Recupero, D., Cambria, E.: Eswc'14 challenge on concept-level sentiment analysis. Communications in Computer and Information Science **475**, 3–20 (2014). https://doi.org/10.1007/978-3-319-12024-9_1, cited By 23
24. Xin, R.S., Rosen, J., Zaharia, M., Franklin, M.J., Shenker, S., Stoica, I.: Shark: Sql and rich analytics at scale. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of data. pp. 13–24 (2013)