Semantic Data Enrichment for Data Scientists

Matteo Palmonari[1], Dumitru Roman[2], Vincenzo Cutrona[1], Nikolay Nikolov[2], and Aljaž Košmerlj[3]

[1] University of Milan-Bicocca
{matteo.palmonari, vincenzo.cutrona}@disco.unimib.it
[2] SINTEF
{dumitru.roman, nikolay.nikolov}@sintef.no
[3] Jožef Stefan Institute
aljaz.kosmerlj@ijs.si

Abstract. The enrichment of a dataset with information coming from third-party data sources is a common data preparation task in data science. Semantic technologies and linked open data can provide valuable support for this task, paving the way for new data-scientist-friendly tools that may facilitate effort-consuming, difficult and boring data preparation activities. In this tutorial, we will provide the audience with: an explanation of the role that semantics play in data enrichment for data science; a review of advantages and limitations of tools, methodologies and techniques for semantic data enrichment available today; a practical dive into the creation of data transformations for enriching the data and the usage of the enriched data to train predictive models. For the latter, we will use tools that support the interactive specification of data transformations and their scalable execution on large datasets and a use case where digital marketing data are enriched to weather-based predictive modeling.

1 Tutorial objectives and description

Data science is in the hot spot today because data analyses play a fundamental role in scientific studies and business decisions. However, it was estimated that about 80% of the effort in a data science project has to be dedicated to data preparation [4]. A data preparation task that is common in different industry-driven and science-driven application domains is data enrichment. Data enrichment is the task of enriching one main dataset, which describes a phenomenon of interest (usually by means of an initial set of variables), with information from third-party data sources, which contain additional variables to feed an analytical model. By using the enriched data, a data scientist can analyze the phenomenon under observation using a larger number of dimensions or train a predictive model using a richer set of features.

On the one hand, data enrichment for data science provides new high-impact application scenarios for exploiting the vast amount of information available

[4] Gil Press. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says, Forbes, March 2016

in the Linked Open Data (LOD) cloud, either as direct source of additional information, or as a bridge for fetching data from information sources outside the LOD. On the other hand, data enrichment provides several challenges for semantic data integration methodologies, revamping known problems under a different perspective, e.g., taxonomy matching, and prompting genuinely new problems, e.g., interactive data reconciliation at scale.

This tutorial has three main learning objectives: (L1) to provide an in-depth understanding of the role that semantics play in data enrichment for data science; (L2) to review advantages and limitations of tools for semantic enrichment available today and of research work relevant to this specific task; (L3) to provide a practical dive into the creation of semantic data enrichment transformations with Grafterizer and ASIA, two integrated tools that support the interactive specification of these transformations and their scalable execution on large datasets, as well as into the usage of the enriched data to train predictive models with the QMiner tool.

In relation to L1, we will use examples and use cases from data science projects developed in two industry-driven H2020 innovation projects, EW-Shopp (www.ew-shopp.eu) and euBusinessGraph (www.eubusinessgraph.eu), which target the development of data-driven innovative services in domains such as digital marketing, eCommerce, and business reporting. In relation to L2, we will discuss the limitations of the approaches based on ad-hoc coding and other semantic agnostic data transformation technologies to deliver data-scientist-friendly solutions. In addition we will discuss research work and prototypes developed in the semantic web, which addressed *some* of the challenges that a data-scientist-friendly solution should solve, and their current limitations (e.g., scalability, support for interaction, etc.). In relation to L3, we will provide an hands-on session to enrich (anonimized) data from a digital marketing agency, which wants to estimate the impact of weather on the impressions of the ads it manages in its campaigns. We will show how to enrich the source datasets with spatial coordinates extracted from Geonames, our "bridge", to eventually fetch data from a weather API. Finally, we will show how to use the enriched data to train a predictive model to estimate the impact of weather on the performance campaign.

Length and draft schedule. The tutorial will be a half-day tutorial, requiring approximately three hours including presentations and a hands-on section. The draft schedule proposed for the tutorial is the following:

- **Data enrichment for data science.** Topics: challenges; semantics technologies as key enablers; analytics with enriched data: use cases from industry-driven data science projects. *Duration: approx. 45 minutes.*
- **State-of-the-art: semantic data enrichment for data science.** Topics: semantic table annotation tools and techniques; doing it with OpenRefine: lessons learned and limitations; data enrichment with batch pipelines: lessons learned and limitations; user-driven data enrichment at scale: the Grafterizer/ASIA approach; QMiner and data analytics on top of enriched data. *Duration: approx. 45 minutes.*

- **A toolkit for user-driven data enrichment at scale endowed with robust analytical modeling.** Topics: the Grafterizer/ASIA data transformation tool: cleaning, annotating and enriching data; QMiner: data analytics on top of enriched data. Approx. 30 minutes.
- **Hands-on session: digital marketing data enrichment with weather data and analytical modelling.** Environment set up; design of the data transformation workflows with Grafterizer/ASIA; batch execution of the transformations on a larger data set; training the predictive model with QMiner; deployment of the predictive model. *Duration: approx. 60 minutes.*

Prior art. The tutorial is presented for the first time. No overlap with previous tutorials exists.

2 Target audience

We expect to attract as audience three different profiles of stakeholders: i) researchers that have expertise in semantic technologies, and, in particular in their application to data integration problems (e.g., ontology matching, semantic reconciliation, table annotation), who will discover new industry-driven, real-world application scenarios for their research results; ii) researchers that have recently entered into the semantic web community, e.g., PhD students and postdocs, who will learn not only about new application scenarios, but also about challenging problems whose solution may have significant impact both in the industry as well as in other science fields, where analyses on top of semantically enriched data may be bring added value; iii) practitioners and workers in the industry field, who will learn about concrete use case for using semantic technology to support data-driven innovation at scale.

To engage audience at the conference we will announce the tutorial through several channels using accounts from our research institutions and from the euBusinessGraph and EW-Shopp innovation projects. In addition to that, we will get endorsements from the projects' business partners and their communication channels. The material will be published online to foster attention by those who could not attend the tutorial, with particular reference to practitioners and industry employees.

3 Speakers' profiles

Matteo Palmonari (PhD) is associate professor at University of Milan-Bicocca. His research interests concern semantics in data integration and artificial intelligence applications. He has been active in the ontology matching field, with a particular focus on interactive matching methods. He is general coordinator of EW-Shopp (Supporting Event and Weather-based Marketing Along the Customer Journey) and unit coordinator in euBusinessGraph, two H2020 projects where the concept of data enrichment is used to support data-driven innovation in the industry. Previously, he has been involved in several other European,

Italian and US projects on semantic interoperability and data integration (COMSODE, SEEMP, SMART, CyberSEES).
Dumitru Roman (PhD) works as a Senior Research Scientist at SINTEF (Norway) the largest independent research organization in Scandinavia. He has wide experience with initiating, leading, and carrying out data-driven and research-intensive projects, participating in dozens of large international projects during the past 14 years in which he has collaborated with various private companies, public sector organizations, universities, and research institutes. He is currently active in the data management field, focusing on innovation projects enabling data-driven business products and services. He has a PhD from University of Innsbruck (Austria) and currently holds an adjunct associate professorship at the University of Oslo (Norway).
Vincenzo Cutrona received the Masters Degree in Computer Science from the University of Milano - Bicocca in 2016 and he is currently a Ph.D. student at the same University, since 2017. His current research activities focus on data integration and semantic enrichment for large-scale data analytics. He has experience in software development and he also worked with a few national companies as a consultant. In the last year, he was involved in an H2020 project as the main maintainer of ASIA, the semantic enrichment component in the EW-Shopp platform.
Nikolay Nikolov is a Research Scientist at SINTEF. He holds a joint Erasmus Mundus M.Sc. degree in Service Engineering from Stuttgart University, University of Crete and Tilburg University from 2014. He has been doing research and software development related to model-driven engineering in the context of cloud computing, big data and the Semantic Web since 2014. During the last four years, he has been involved in several European Union FP7 and Horizon 2020 projects with a focus on design and adaptive deployment of applications in the cloud, and linked open data integration (PaaSage, MODAClouds, DaPaaS, proDataMarket). Nikolay Nikolov leads the development of the DataGraft platform, which is used as a core component in the EW-Shopp platform.
Alja Komerlj (PhD) is a researcher in the Artificial Intelligence Laboratory at the Joef Stefan Institute - the largest research institute in Slovenia. His background is in artificial intelligence and machine learning with his current work mostly focusing on text mining and natural language processing. He has extensive experience from work on research projects from both the European (FP7, H2020) and national levels as well as commercial projects for international clients (Bloomberg L.P.).

Acknowledgements

This tutorial is based on research carried out with funding from the European Unions Horizon 2020 research and innovation program under grant agreements No 732003 and No 732590.