

# Semantically-enabled Optimization of Digital Marketing Campaigns\*

Vincenzo Cutrona<sup>1</sup>, Flavio De Paoli<sup>1</sup>, Aljaž Košmerlj<sup>2</sup>, Nikolay Nikolov<sup>3</sup>,  
Matteo Palmonari<sup>1</sup>, Fernando Perales<sup>4</sup>, and Dumitru Roman<sup>3</sup>

<sup>1</sup> University of Milan-Bicocca, Milano, Italy - *{lastname}@disco.unimib.it*

<sup>2</sup> JSL, Ljubljana, Slovenia - *aljaz.kosmerlj@ijs.si*

<sup>3</sup> SINTEF AS, Oslo, Norway - *{firstname.lastname}@sintef.no*

<sup>4</sup> JOT Internet Media, Madrid, Spain - *fernando.perales@jot-im.com*

**Abstract.** Digital marketing is a domain where data analytics are a key factor to gaining competitive advantage and return of investment for companies running and monetizing digital marketing campaigns on, e.g., search engines and social media. In this paper, we propose an end-to-end approach to enrich marketing campaigns performance data with third-party event data (e.g., weather events data) and to analyze the enriched data in order to predict the effect of such events on campaigns' performance, with the final goal of enabling advanced optimization of the impact of digital marketing campaigns. The use of semantic technologies is central to the proposed approach: event data are made available in a format more amenable to enrichment and analytics, and the actual data enrichment technique is based on semantic data reconciliation. The enriched data are represented as Linked Data and managed in a NoSQL database to enable processing of large amounts of data. We report on the development of a pilot to build a weather-aware digital marketing campaign scheduler for JOT Internet Media—a world leading company in the digital marketing domain that has amassed a huge amount of data on campaigns performance over the years—which predicts the best date and region to launch a marketing campaign within a seven-day timespan. Additionally, we discuss benefits and limitations of applying semantic technologies to deliver better optimization strategies and competitive advantage.

**Keywords:** Semantic enrichment · Big data analytics · Digital marketing.

## 1 Introduction

Digital Marketing is a growing industry<sup>5</sup> where the engagement strategy has evolved from non-personalized to highly personalized user targeting, with the

---

\* The work in this paper is partly funded by the EC H2020 projects EW-Shopp (732590) and euBusinessGraph (732003). Authors are listed in alphabetical order. Corresponding authors: Dumitru Roman and Matteo Palmonari.

<sup>5</sup> Revenue in the Digital Advertising market amounts to US \$63,469m in 2019, according to <https://www.statista.com/outlook/216/100/digital-advertising/worldwide>.

aim of reaching the target users with relevant content and promoting user actions (clicks, sales, leads, etc.) on service delivery platforms. In this strategy shift, marketing agencies and marketing departments of large companies need to collect and process large amounts of marketing-related data, and implement new management strategies to optimize the use of marketing budgets in the most profitable campaigns. This requires technical knowledge for managing data for service delivery (e.g., campaigns optimization, programmatic buying, content generation and delivery, organic positioning) and analytical skills to process the existing raw data, analyse actions performance, and generate insights motivating future marketing strategies. Such knowledge and skills are often not available to digital marketing agencies, especially small and medium in size.

In this paper we report on a pilot implementation that uses cutting-edge data semantics and analytics technologies in the context of the rather technology-conservative digital marketing domain. The pilot was facilitated by JOT Internet Media<sup>6</sup> – a Spanish SME operating in the digital marketing domain. JOT is specialized in Web traffic generation by means of investing in sponsored ads in the main search and display platforms (e.g., Google, Bing, Facebook). The massive implementation of its digital marketing campaigns enables daily collection of huge amounts of data related to campaigns’ performance. Performance indicators (clicks, impressions, CTR, location, date, time, identifiers, keywords, device, ad platform, etc.) are collected and analyzed daily. Currently, this activity is based on account managers’ experience to optimize both the campaign management and bidding strategy to engage the audience and generate actions in service delivery platforms’ landing pages, which generates JOT’s revenue streams. JOT aims to create a new data-driven campaign management service based on the integration and enrichment of its performance datasets with weather forecasts to predict campaign impact and optimize the budget distribution in marketing campaigns.

Semantic technologies (for enriching the data) and machine learning (for analytics of the enriched data) were identified as promising technologies by JOT to support the creation of its new data monetization service. Together with R&D and technology providers, JOT created a pilot to assess benefits and limitations of semantic technologies and machine learning. In this paper we present and discuss experiences in the design and implementation of this pilot. Contributions of this paper include the definition of a pilot for the use of semantic technologies in the digital marketing domain, a generic approach for marketing campaigns performance data enrichment and analytics, as well as an implementation of the approach using cutting-edge tools, together with experimental insights.

The rest of this paper is organized as follows. Section 2 provides the necessary background for the developed pilot. Section 3 presents the pilot together with its requirements. Section 4 outlines the developed approach for semantic enrichment and data analytics and reports on experimental insights. Section 5 presents related work and discusses the advantages and limitations of the used

---

<sup>6</sup> <https://www.jot-im.com>

technologies and the overall approach. Finally, Section 6 summarizes the paper and outlines avenues for potential future work.

## 2 Background and motivation

JOT’s digital marketing focus is on advertising platforms such as Google Ads<sup>7</sup>, where advertisements are placed using Real-time Bidding (RTB). As background for understanding the developed pilot, we give a brief overview of how digital marketing campaigns are executed on such platforms and discuss the opportunities and motivation for semantic data enrichment and analytics in this context.

Upon a user search, platforms such as Google Ads run a bid, where different marketing campaigns compete to display an ad (e.g., an ad linking to a landing Web page). A digital marketing campaign defines, in principle, a set of keywords and for each keyword the maximum cost per click (MaxCPC) paid in a bid for that keyword. The term *impressions* for a keyword refers to the number of times some ad has been displayed on the sponsored ads space in Google Search when users search for that keyword and a campaign wins a bid on that keyword. Therefore, a keyword can generate (and be associated with) impressions only if at least one digital marketing campaign is active for that keyword, and someone searches for that keyword. The number of impressions is considered a key performance indicator (the higher the better), hence a primary goal for a campaign is maximizing the amount of impressions. In this way, the advertised landing page has potentially more visitors and opportunities to increase its brand awareness and sales, which is the ultimate goal of a campaign. For each day, impressions can be counted and visualized to let strategy experts evaluate the performance and define the strategy for the next campaigns.

Performance indicators are influenced by a variety of factors, some related to the marketing domain and campaign implementation (e.g., maximum bid, number of competitors and degree of matching between the landing page and user search), and some related to external factors which are out of control by the companies running the campaign. For example, JOT found evidence that weather events can affect the performance of campaigns in a sensible way. Exploratory analyses found examples where the number of impressions of some specific keywords or keyword categories showed abnormal increase due to the weather. For example, analyzing data from February 2016 revealed that depending on the day and the rain forecast, population in the Madrid region had more interest in “burger at home” or in keywords related to the “DiningNightLife” category. This effect can be replicated to other less obvious keywords, making it extremely useful for marketers and accounts to adjust campaign launch and bidding strategy so they can optimize budget consumption and increase their impact in terms of impressions and ad clicks.

The enrichment of digital marketing campaigns with third-party event-related data and their subsequent analysis can provide several benefits to companies running such campaigns:

---

<sup>7</sup> <https://adwords.google.com>

**Advanced data insights:** i) correlation between marketing performance indicators and external variables such as temperature, probability of rain, light hours; ii) identification of new trends and patterns; iii) information useful for bid adjustment for the affected keywords.

**New services for campaign scheduling:** i) define campaign launch scheduling according to the influence of external factors to optimize the impact; ii) consultancy services, such as the identification of key keywords enabling higher impact depending on external factors; iii) evaluation of impact depending on campaign properties (country, topic, and timing).

### 3 Pilot: The weather-aware campaign scheduling service

The developed pilot – a weather-based campaign scheduling service – was motivated by marketing experts observations about the weather-sensitiveness of certain keywords. This triggered the need for a systematic approach to detect dependencies between weather variables and ads impressions for certain keywords. The assumption was that if a keyword depends on a set of weather variables, then a predictive model to estimate the number of impressions (or at least peaks in impressions) can be defined. Using such a model, a new service can be devised to collect weather forecasts for the geographical regions involved in a given campaign for seven days and to estimate the best date to launch the campaign in each region in that time frame. The service is meant to be used by campaign managers at JOT to schedule marketing campaigns.

The data processing workflow designed as part of this service is depicted in Figure 1 and explained as follows.

**Performance data time-series collection.** Time series about campaign performance are collected as historical data aggregated by location (to support weather forecasts) for a period of time long enough to train a predictive model; an example is the top-left-hand table in Figure 1 (*#im* stands for number of impressions).

**Enrichment of campaign performance data.** Performance data time-series are enriched with weather variables relevant to train predictive models; an example of enriched table is the top-right-hand one in Figure 1 (*C* represents temperature in Celsius; *mm* precipitation in millimeters; and *x/+y* forecasts for the day *x* plus *y* days).

**Analysis of enriched data.** Enriched performance time series are used to build predictive models that, based on location and a 7-day weather forecast, can estimate the number of expected daily impressions; notice that the model may not be developed for each keyword, e.g., when accuracy of prediction is too low to provide a reliable model.

**Run-time execution of predictive models.** When we want to activate a campaign for a keyword and a model exist, that model can be applied using weather forecasts to return a weather-based schedule to campaign managers (as shown in the dashed line box at the bottom of Figure 1).

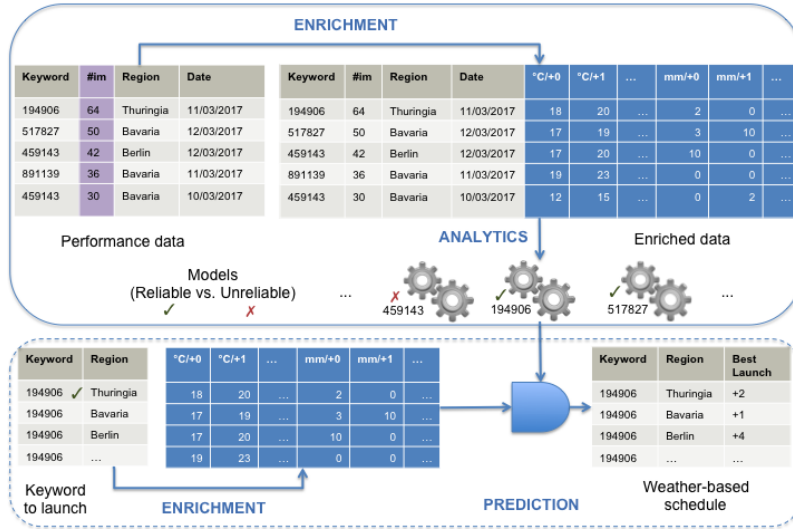


Fig. 1. Data processing for weather-based campaign scheduling.

Two main **data sources** were identified as relevant for the pilot, as follows.

*Marketing campaign statistics from JOT.* For the pilot, we consider campaigns run in Germany and Spain for a total of  $\sim 22$  million keywords, which are associated with (all) Google categories and span over 2016 and 2017. Row data are associated with spatial references based on Google GeoTargets (*location identifiers used by Google*), at city and region level. Data were provisioned in CSV format ( $\sim 100$ GB and  $\sim 500$ M rows). The dataset includes 21 columns that cover: keyword ids (unique identifiers of the keywords in Google), keywords (the keyword tokens), several variables describing *matches* for the ads, i.e., indicators of their performance (e.g., clicks, impressions, ad position) measured in a specific location (a city and its region), specification of campaign data, (country, language, category, listing and match type), and a category associated with the keyword (first level category in the Google taxonomy).

*Weather forecast data from ECMWF.* Data provided via APIs by the European Centre for Medium-Range Weather Forecasts (ECMWF)<sup>8</sup>, amounting to 85PB in GRIB<sup>9</sup> format. The GRIB format represents weather data on a grid, where intersection points are specific coordinate pairs. Given the coordinates of a city, the city-level weather data are computed by interpolating the information available for the nearest four points in the grid.

Based on the envisioned data workflow and the identified data sources, a number of **challenges** were identified, as follows.

<sup>8</sup> Meteorological Archival and Retrieval System.

<https://software.ecmwf.int/wiki/display/UDOC/MARS+user+documentation>

<sup>9</sup> General Regularly-distributed Information in Binary form.

<http://www.wmo.int/pages/prog/www/DPS/FM92-GRIB2-11-2003.pdf>

*Data enrichment.* Weather forecast data are in GRIB format, which is very space-efficient, but not time-efficient to support queries due to its binary nature (Figure 1 reports weather data in an intuitive format to simplify the example). Weather data can be queried using coordinates (longitude and latitude), which are not present in the JOT dataset or in Google location data. Enriching each row with coordinates is a prerequisite to fetch weather data, hence location toponyms in Google need to be reconciled with a geospatial knowledge base where locations have such coordinates in their descriptions. Moreover, data managers at JOT are used to work with tabular data and prefer to design these transformations (reconciliation and extension) leveraging a tabular view over their data.

*Data analytics.* Performance data provide signals that are often scattered (because a keyword may be active for a limited number of weeks or days), weak (because of few impressions) and noisy (because of other parameters may affect the campaign performance, e.g., competing bids over a time span). Thus, building models that are accurate-enough to be exploited in a production service is difficult and not possible for a large number of keywords.

*Scalability.* The size of the datasets is a challenge, even the analysis of campaigns for a single country requires managing roughly  $\sim 1$ TB of data. JOT runs campaigns targeted to more than 70 countries worldwide, hence the size of the data in the scope of the analysis may become huge, which means that enrichment should be based on flexible technologies to cover more countries upon request (e.g., minimal changes on the schema of the data should be required when adding countries, and coordinates should be fetched only for the new locations). Therefore, simplifying and making the enrichment tasks more scalable is a key requirement to support the analysis of campaign performance data at full scale.

## 4 Approach for semantic enrichment and analytics

In order to address the main challenges identified above, we devised an approach guided by the following principles.

**UI-based design of data transformations.** Data transformations should be designed using a UI that supports users in establishing how to transform the original data (i.e., tables of data they are familiar with), and in displaying the results in an understandable way (i.e., readable by non-experts in semantics). Moreover, since datasets can be huge, we need to adopt a strategy of working on samples to manage Big Data.

**Batch execution of data transformations.** The transformations resulting from the UI-based design will be applied to the whole dataset using a scalable platform in batch mode.

**Transparent use of semantic technologies.** Data enrichment can be broken down into two tasks: *data reconciliation*, where identifiers in the source data are reconciled against a *knowledge base* (KB), and *data extension*, where the identifiers in the KB are used to fetch additional data from third-party

sources. Semantics, KBs and Linked Data, are used to support these tasks in a way that is as much as possible transparent to generic users.

**Replicability and adaptation of transformations.** The transformations resulting from the UI-based design should be repeatable to new datasets that hold the same structure and content (e.g., performance of the campaign in the same country for a different time period), and adaptable (e.g., performance of a campaign with different countries, which may require just an adaptation of the configuration for the reconciliation transformation).


Following these principles our approach is composed of the following steps.

**UI-based data transformation design.** The user uploads a data sample and designs transformations to clean the data (e.g., date formatting) after which she enriches them with third-party data. Working on a sample, the user can immediately view the effect of the transformations to tune them.

*Knowledge bases* (KBs) are used to bridge across different systems of identifiers used in the corporate and in the third-party data sources. An example is GeoNames, which provides a convenient reference KB in our case due to complete coverage, multilingualism and information quality. However, other cross-domain KBs such as WikiData and DBpedia may be useful to access other kinds of information associated with locations.

The user reconciles the values, e.g., spatial references, against shared KBs by using *reconciliation services* for these KBs from the UI. By using a reconciliation service on the sample data the user configures the reconciliation service in such a way that it can be applied later on top of the unseen data processed in batch mode. An example of this configuration is setting a similarity threshold for the algorithm after having explored its impact on the data. Once values are reconciled on the sample data, the user can use a *data extension service* provided by a third-party source, e.g., GeoNames or a weather API, to specify the values to add to each row. This specification includes setting the join conditions on single or multiple values in a row from a widget. For example, given the GeoNames identifier of a location, longitude and latitude can be fetched from GeoNames; given longitude, latitude and a date, weather variables can be fetched from a weather data source. Similarly, if the user needs to normalize impressions by the population density of the region where they have been measured, the population can be fetched from GeoNames, while area can be fetched from Wikidata after reconciling the GeoNames identifiers to Wikidata identifiers (using *same-as* links if available, or full-fledged reconciliation). Multiple reconciliation and extension steps can be applied to the data, as depicted in Figure 2.

Once the table has been extended, data can still be used in tabular format or *mapped to a graph schema*, e.g., an ontology, using schema-level enrichment services from the UI. During the enrichment process, algorithms similar to the ones developed for automatic *semantic table annotation* [10] support the user by providing suggestions on reconciliation and mapping to a schema. Data enrichment is in fact an interesting new application field for semantic table annotation approaches with human-in-the-loop. All the transformations designed by the user with the UI are transformed into code that can be executed in batch mode.



Keyword	#im	City	Region	ID (Geonames)	Latitude (Geonames)	Longitude (Geonames)	ID (Wikidata)	Area (Wikidata)	Temp (ECMWF)	Date
194906	64	Altenburg	Thuringia	2622542	50.98763	12.43684	Q1205	45.6 km <sup>2</sup>	18°	11/03/2017
517827	50	Inglostadt	Bavaria	2951839	48.76508	11.42372	Q980	133.35 km <sup>2</sup>	17°	12/03/2017
459143	42	Berlin	Berlin	2950157	52.52437	13.41053	Q648102	891.68 km <sup>2</sup>	17°	12/03/2017
891139	36	Munich	Bavaria	2951839	48.13743	11.57549	Q980	310.71 km <sup>2</sup>	19°	11/03/2017
459143	30	Nuremberg	Bavaria	2951839	49.45421	11.07752	Q980	186.45 km <sup>2</sup>	12°	10/03/2017

Fig. 2. An example of data enrichment.

**Data enrichment: Batch execution of data transformations on a scalable infrastructure.** Data transformations are executed in batch mode in order to support the enrichment of data that are too big to be controlled interactively from a UI. For very large datasets, like JOT’s data about campaign performance, using a single host is not sufficient. We employ a scalable infrastructure that can be operated on the cloud and scaled based on performance needs and available budget.

**Analytics: Building the models.** Once data are enriched, a data analyst will define useful aggregation functions over the data and test training of different models using standard machine learning methodology. The level of aggregation is often defined by the data analyst also based on an evaluation of the accuracy of the predictions for different levels of aggregation (e.g., signals of impressions at the city level may turn to be too weak to be used, so that data need to be aggregated at the region level).

**Preliminary data pre-processing to meet performance constraints.** While we could apply the above mentioned tool-supported approach to the input data, the large amount of data to be processed, the tight requirements on the reconciliation steps, and the scattered distribution of performance data required the following data processing steps before the design of the data transformation pipelines:

- Pre-processing of third-party data downloaded in advance (e.g., weather forecasts for a country considered in the analysis);
- Data linking between systems of identifiers used in the corporate data source (e.g., Google GeoTargets and GeoNames);
- Filtering of campaign performance data based on temporal continuity and signal strength; because signals provided by performance indicators may be too discontinuous or too weak to train a reliable model, performance data should be filtered by strength and continuity in order to analyze the performance of keywords where patterns are more likely to be found.

We built and extended a set of software components to implement the above described approach. The components are organized in a toolkit depicted in Figure 3, referred to as the EW-Shopp toolkit, to support event-based data enrichment and analytics across a number of related business cases.



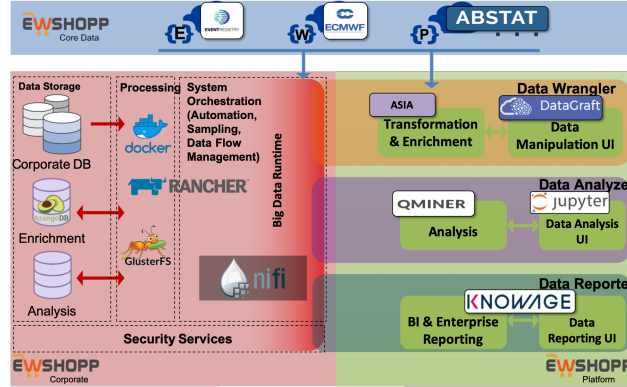


Fig. 3. EW-Shopp toolkit.

The toolkit consists of: *the data wrangler layer*, with **Grafterizer** and **ASIA**, two tightly integrated components for the design and execution of data cleaning and enrichment transformations in DataGraft [11]; *the data analyzer layer* with **QMiner**, one library for efficient data analytics, with a set of scripts to support weather and event-based analyses; *the reporting layer* with **Knowage**, a tool for data visualization whose use is not discussed in this paper. In addition, **APIs** are used to simplify the access to third-party data, including, e.g., the ECMWF weather data source, the Event Registry<sup>10</sup> (the usage of which is not discussed in this paper) and ABSTAT (a knowledge graph profiling tool that is used by ASIA). Grafterizer and ASIA are deployed on the cloud. Data transformations as well as Grafterizer and ASIA backends are enclosed into containers to be run on the company’s private cloud infrastructure, where they can be executed using a distributed **Big Data infrastructure** that supports parallel execution. We briefly describe the main features of Grafterizer, ASIA, QMiner and the deployment of the Big Data infrastructure, before describing the processing steps applied to implement the pilot.

**Grafterizer** [13] is a tool that supports the design of data transformations to clean and manipulate tabular data (including transformation to RDF) through a UI. In addition, it provides profiling and quality assessment features to support the process. Grafterizer can be used to produce data that can be onboarded on top of different databases. To develop data transformations for large-scale digital marketing data it has been modified to generate imports to ArangoDB, a multi-model database, which allows the manipulation of large data sources with its support for graph, document and key-value storage/querying capabilities. Grafterizer’s transformations are encoded in Clojure scripts in such a way that its backend can execute the data transformations in batch mode. Data transformations can be saved and replicated on new data sources. The main challenge

<sup>10</sup> <http://www.eventregistry.org>

addressed to support transformations on large-scale data was the deployment of the transformations on a Big Data infrastructure that supports parallelization.

**ASIA**<sup>11</sup> (Assisted Semantic Interpretation and Annotation of tables) is a new tool to help users annotate and enrich a table using semantics in the process. ASIA supports schema-level annotations to map the table schema to existing vocabularies and ontologies by using vocabulary suggestions powered by the ABSTAT profiling tool [12], and instance-level annotations by using reconciliation services. Finally, it supports data extension services to enrich a table with third-party data sources. The ASIA front-end supports these features through a UI encapsulated in the Grafterizer tool. The transformations implicitly encoded into users' annotations can be also replicated and natively executed by ASIA's backend in batch mode on large datasets. To support enrichment of large scale data we addressed the challenge of making data reconciliation and extension services executable in a scalable and efficient manner on a parallel architecture deployed in the Cloud via containerization. The latter is a novel feature compared to existing tools that provide some support for enrichment-related tasks.

**QMiner**<sup>12</sup> [4] provides fast modeling and execution of analytics on large-scale data providing a large number of machine learning techniques. It is designed to efficiently process both structured and unstructured data, storing and indexing, in a way that makes machine learning algorithms scalable. The algorithms themselves are implemented in a C++ library, which is wrapped in a JavaScript API for ease of use and flexibility, making the deployment of the models in production environments simple without sacrificing performance.

The **Big Data infrastructure** for the scalable execution of data transformations – including semantic enrichment – on the cloud is designed to be deployable on heterogeneous infrastructures that may be managed by a platform user. It consists of a container system, a container orchestration system and a distributed file system. The container system (Docker) is used to encapsulate data workflow steps, which are used to pre-process, transform, semantically enrich, and onboard data. The container orchestration system (Rancher<sup>13</sup>) is used to manage and scale up the data workflows by enabling their deployment across a managed set of hosts provisioned for data workflow execution. The distributed file system (GlusterFS<sup>14</sup>) is the data communication medium through which data are passed between steps and is also used for storage of intermediate results during processing.

The above discussed toolkit has been applied to support the end-to-end data processing as following.

**Data enrichment.** Data enrichment includes the following steps.

*Data ingestion.* For the pilot, we chose the country of Germany and data about keyword activity during 2017 for four high-level categories: Business,

<sup>11</sup> <http://inside.disco.unimib.it/index.php/asia>

<sup>12</sup> <http://qminer.ijs.si>

<sup>13</sup> <https://rancher.com>

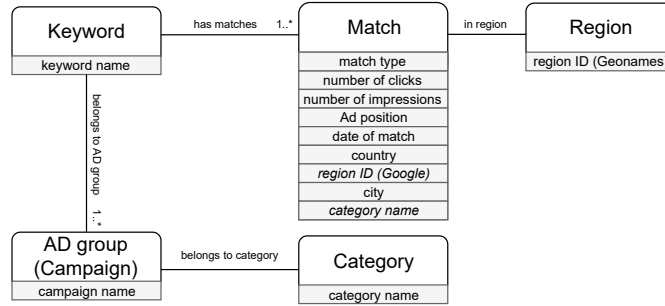
<sup>14</sup> <https://www.gluster.org>

Travel, Health, and Vehicles. Data are structured as described in Section 3. The dataset consists of a total of 15 million keywords with all of their matches during 2017 (matches make up most of the data). Data are uploaded to an FTP server (hosted on Amazon Cloud). Weather data are collected from ECMWF using a new weather API (the one depicted in Figure 3), which was built to support time-efficient queries on the data, solving the problems caused by the space-efficient GRIB format used in the original APIs. Seven days weather forecasts as predicted daily along 2017 for the whole Germany are downloaded in advance. The data are also further processed to make enrichment more efficient at large scale, in particular, data are transformed into JSON using Grafterizer and uploaded to ArangoDB. This step implements pre-fetching of third-party data mentioned in Section 4 but also adds an additional processing step using a library that complements the new weather API. Weather forecasts are organized into documents where each document presents a weather forecast made at a given date (the one to be matched against the match date), a given region, and a given offset (e.g.,  $x/+y/+z$ , where  $x$  represents the date,  $+y$  represents the day offset, and  $+z$  represents the hour offset). The region is specified by its GeoNames identifier, after interpolating the grid points in the raw data using region bounding boxes provided in GeoNames. This semantic-enriched graph-based representation of weather data supports more efficient enrichment in batch mode.

*UI-based data transformation design.* A sample is extracted from the ingested data and uploaded to Grafterizer. Data transformations are designed using the UI of Grafterizer and ASIA, the first one to clean the data and a reconciliation service of the second one to reconcile Google GeoTarget spatial identifiers against GeoNames. Thanks to the semantic-enriched graph-based representation of weather data the user does not need to add longitude and latitude to the data and can proceed to specify the desired weather variables and the columns used to join data using a weather extension widget in ASIA. For UI-based reconciliation and extension specification in ASIA we refer to online documentation<sup>15</sup>.

Data are then mapped to a graph model using Grafterizer’s UI to support their transformation in JSON and upload into ArangoDB. In this process, identifiers are associated with performance statistics using triples consisting of the region, the date (of forecast) and the offset, similarly to the identifiers associated with weather documents. In this way it is possible to naturally join performance data and weather data upon upload in the database. The specification of the transformations is extracted as Clojure code. Although ASIA supports interactive reconciliation, in order to make the reconciliation process faster and more reliable we computed links between Google GeoTargets in Spain and Germany and GeoNames using the Silk data linking tool [9] and manually validated them. As a result, the resulting reconciliation service used a identifier-to-identifier dummy reconciliation system. Caching and other strategies are used in batch enrichment to speed up the process (a detailed explanation of these strategies is beyond the scope of this paper).

<sup>15</sup> See video at <https://youtu.be/4amLd4biYcs> and the *Semantic Data Enrichment for Data Scientists* tutorial at <https://ew-shopp.github.io/eswc2019-tutorial>.



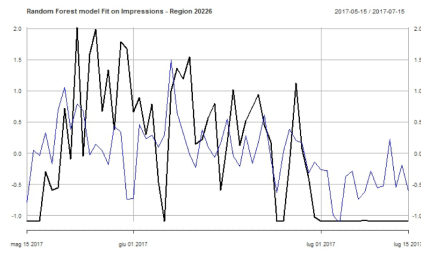
**Fig. 4.** The Pilot data schema.

*Batch execution of data transformations on the Big Data infrastructure.* Closure code and ASIA backend are packaged into containers and executed on the scalable Big Data infrastructure. Enriched performance data are uploaded to ArangoDB (according with a schema depicted in Figure 4), where they are integrated with weather data through shared identifiers generated during the transformation to the graph model. From this step on, data are available to the data analyst. The graph database here reduces disk usage and the amount of data generated in the transformation by avoiding redundant information that would be contained into a big integrated table. Our implementation can be easily adapted to work only with tabular data, if a user prefers to use a redundant tabular data model instead of a graph database. We tested different configurations and evaluated their impact on the scalability of the solution.

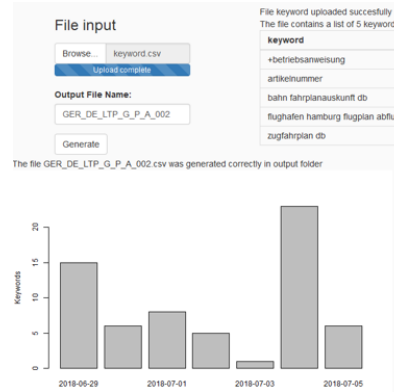
**Data analytics.** The goal of the analytics is to build a model that predicts keyword behaviour from weather data. If the model predicts a spike in the level of impressions of a keyword, that is a good time to run a campaign using that keyword.

For modelling, all impressions were aggregated at the region level. As mentioned before, useful keyword impression data is limited to times when the keyword was part of an active campaign. As the information on when the individual keywords are active was not available, this selection process had to be performed manually. As this limits the capacity of the analytics considerably, a better solution is considered as part of future work (see discussion in Section 5). The keywords were ranked by the overall volume of impressions and the ones with the highest volume were selected for analysis. The impressions were standardized by region to make them comparable. One region was excluded as the validation sample, the rest were split randomly into a training and test set (70% - 30%). To simplify analysis we ignored all weather features with zero variance. In the pilot we applied the Random Forest [1] model which outperformed other models in our tests. The model returns the number of impressions expected in a day. We used Root Mean Square Error (RMSE) to evaluate the goodness of fit of the model.

As explained before, we were only able to run the analytics testing for a limited number of keywords due to manual work in the process. Here we report the results for a single keyword, “deutsche bahn fahrplanauskunft” (“German railways timetable” in English), as an illustrative example. We trained the model over the period from 15.5.2017 to 15.7.2017 where the keyword had the most activity. The Random Forest model achieved RMSE of 0.77 on the test set and RMSE of 0.87 on the validation set. Figure 5 shows the predicted values and the actual values for the validation set (Region 20226 in Google GeoTargets)<sup>16</sup>.



**Fig. 5.** Actual (black) and predicted (blue) values for standardized number of impressions in a region.



**Fig. 6.** Service interface.

The result of the analyses conducted on  $\sim 40$  keywords were provided to the weather-based campaign scheduler service (the interface of which is depicted in Figure 6) that is offered to campaign managers.

## 5 Related work and discussions

From JOTs commercial perspective, there is no similar solution on the market enabling the access to a predictive model identifying the most affected keywords and the time when it is more profitable to launch a campaign.

Our review of related work shows that the field of digital marketing has been aware of semantic technologies for some time, with books in the field dedicating entire chapters to the topic of semantic Web [5, 14]. However, semantics are mostly used for content organization with the goal of optimization of website ranking on search engines, though marketing research recognizes their wider

<sup>16</sup> The line comparison in Figure 5 shows a comparison of the actual and predicted level of impressions for an anecdotal example. Its purpose is more illustrative as it does not reflect global performance of the approach, though it does suggest what level of possible deviation a marketing professional has to take into account when using the model.

potential [3]. According to our review, most of the applications of semantic technologies in digital marketing have been for user profile modelling. In [7] and [8] an approach is described for using semantic annotations of users' browsing behaviour (i.e., the content they interact with) to model them and predict the best marketing content for their preferences. Similar approaches have been proposed for support of organization and management of user interactions in social media, such as the SEMO platform for customer social network analysis based on semantics and emotion mining [6].

We support the data enrichment process with interactive semantic table annotation techniques. Several methods have been proposed to automatically interpret the semantics of tables but they have been usually tailored to many but small tables. For these methods we refer to related work reported in a recent paper [2], comparing our work to (the few) full-fledged interactive table annotation tools. Karma<sup>17</sup> [10] and Odalic<sup>18</sup> provide sophisticated schema-level annotation suggestions, but, to the best of our knowledge, they lack UI-powered services for the reconciliation and extension of column values. OpenRefine has been an important source of inspiration, as it provides a neat user interface and enrichment services, which are used by a relatively large community of end users<sup>19</sup>. We aim at generalizing and improving on the OpenRefine data enrichment features (e.g., more extension services than Wikidata and replicability of transformations) and make them applicable to large amount of data by executing reconciliation and extension services on a scalable infrastructure (batch versions of OpenRefine seem not to support enrichment features, which call external services).

Tools such as Silk<sup>20</sup> and LIMES<sup>21</sup> provide capabilities for data interlinking, however such tools do not perform linking of values occurring in tables. A user would need to use such interlinking tools first (which requires good knowledge of RDF) and then would need to upload links to a triple store and do enrichment via SPARQL queries, switching from one tool to another. Here we support linking within the tabular manipulation tool itself. Furthermore, no approach has worked on linking Google GeoTargets, a crucial dataset to work with in the digital marketing domain (used also in GoogleAnalytics). The novelty in this context is the relevant, high-impact problem domain and the semantics-based solution we devised to successfully address the problem.

**Discussion on the relevance of the results to JOT.** The enrichment pipeline was used to process data for 2016 and 2017 for ~22 million keywords from ~47 thousand campaigns and is now in use to collect and enrich data for 2018 and 2019 analyses. The enriched data were handed over to a data scientist who filtered the most promising keywords to produce predictions to be evaluated by JOT experts. The predictions were tested on a smaller number of keywords, with best results judged valuable for usage in production by the JOT team.

<sup>17</sup> <http://usc-isi-i2.github.io/karma>

<sup>18</sup> <https://www.adequate.at/odalic>

<sup>19</sup> <http://openrefine.org/my%20category/2018/07/16/2018-survey-results.html>

<sup>20</sup> <http://silkframework.org/>

<sup>21</sup> <http://aksw.org/Projects/LIMES.html>

RMSE scores were judged valuable for usage in production by JOT experts for the following reasons: (1) the prediction is used to find peaks to determine the launch date (limited accuracy in curve prediction is acceptable for monetization purposes, which is eventually ground on exploiting the peaks); (2) the launch will involve many keywords when the analyses are scaled up. Thus, the JOT team believes that micro-improvements on many individual keywords, even with some errors, will lead to monetization at scale. The pipeline is compliant with JOT cloud infrastructure (that was a big challenge) and is intended for tech staff (not plain marketeers). It is also replicable, as the CSV schema will not change. Thus, the enrichment pipeline is now mature, while the analytics have to be scaled up to go in production.

**Discussion on the use of semantic technologies.** We briefly discuss here some lessons we learned about the role of semantics for enriching data in industry-driven data science projects, discussing (Pros) and (Cons), with the latter referring to limitations and open issues.

**(Pros)** The approach and its implementation described in this paper successfully supported the enrichment of large amount of marketing campaign statistics with a large amount of third-party sources, for subsequent analysis. For a set of manually selected keywords weather-based predictions are in fact reliable and usable to support campaign managers of a digital marketing company. Effective enrichment technology can bring much value in the digital marketing domain, where in-depth analytics are key to success and the variables available in the source data is limited by reporting tools provided off-the-shelf by digital marketing platforms (e.g., Google AdWords). **(Cons)** So far we considered weather-based enrichment and many more challenges are ahead - e.g., understanding which data in the LOD cloud can be useful in this domain, and using media coverage signals, extracted from semantic event engines like the Event Registry. In addition, scaling up the analysis requires overcoming the manual selection of promising keywords and the weakness and discontinuity of performance signals over long periods of time (e.g., impressions). To solve this issue we are currently experimenting keyword clustering methods based on multi-lingual word embeddings.

**(Pros)** Semantics revealed to be a key enabler to support and scale up the enrichment process: reconciliation against reference KBs (e.g., GeoNames) and data interlinking (Google GeoTargets vs. GeoNames) are key pillars for designing enrichment pipelines and for enabling strategies to execute these pipelines in a more efficient way (e.g., by using graph-based databases). **(Cons)** Little work has been done to interlink data, e.g., Google GeoTargets or Google Categories, used in digital marketing platforms that serve millions of companies. Coverage of interlinks between these sources and other sources in the LOD cloud must be improved.

**(Pros)** Semantic enrichment is a promising yet underinvestigated application of semantic table annotation techniques to facilitate a variety of business analytics. Our contribution targeted mainly engineering problems related to their application for large scale enrichment, thus complementing previous work that

focused on intelligent table interpretation. **(Cons)** A better integration of these aspects is a key challenge we are currently addressing.

**(Pros)** Inspired by tools used by a large user base such as OpenRefine, we developed an approach where semantics are used in a way that is maximally transparent to the user, who uses reconciliation and extension as services from a user interface. We have then shown that annotations can be made with transformations that can be executed in settings that meet the key business requirements (scalability, cloud-based deployment). **(Cons)** Some (semantic) pre-processing steps, in particular for weather data, had to be used, which slightly change our vision, for which integration of corporate data with external sources can be solved by applying a sequence of reconciliation and extension steps. To further optimize the enrichment process and validate the table extension approach, we need to better understand trade-offs between using a big denormalized table and using graph-based representations (which we used to limit space usage and disk writing time).

Finally, we mention that weather-based enrichment (with an external large data source) shows that our approach can be applied in complex and large-scale scenarios. However, there are plenty of LOD sources that are underused despite their potential value, also because of users' limited knowledge of their content and of semantic technologies. Building extension services on top of LOD sources is straightforward, but better support for reconciliation against these sources is needed. The availability of tools to support semantic enrichment in business contexts in the era of analytics may also foster the consumption of LOD beyond the semantic Web enthusiasts.

## 6 Summary and outlook

Digital marketing is a domain that has traditionally been rather conservative in adapting new technologies. At the same time, it is moving more and more towards exploitation of data in new ways. With this paper we presented an experiment in using semantic technologies for enriching marketing campaigns data and machine learning to analyze the enriched data, with the final purpose of implementing a new campaign management methodology optimizing the impact of campaigns for a digital marketing company (JOT). In this process, an end-to-end process was devised, from enrichment of data about digital marketing campaigns performance with third-party event data, through the analysis of the enriched information asset using machine learning techniques, to development of value-added services on top of the analytics results. This paper demonstrated the potential use of semantic technologies (with focus on semantic enrichment) for digital marketing — an application domain that has received relatively little attention in the semantic Web community in comparison with other application domains.

As part of future work we consider integrating the pilot into the production systems at JOT and increasing the number of the analytics tasks on the enriched



data. Another direction for future work could be the use of OpenWeatherMap (OWM)<sup>22</sup> data as an alternative to weather data from ECMWF<sup>23</sup>.

## References

1. Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (Oct 2001)
2. Chen, J., Jimenez-Ruiz, E., Horrocks, I., Sutton, C.: Colnet: Embedding the semantics of web tables for column type prediction. In: *AAAI* (2019)
3. Erragcha, N., Romdhane, R.: New faces of marketing in the era of the web: From marketing 1.0 to marketing 3.0. *Journal of Research in Marketing* **2**(2) (2014)
4. Fortuna, B., Rupnik, J., Brank, J., Fortuna, C., Jovanoski, V., Karlovcec, M., Kazic, B., Kenda, K., Leban, G., Muhic, A., Novak, B., Novljan, J., Papler, M., Rei, L., Sovdat, B., Stopar, L., Grobelnik, M., Mladenec, D.: Qminer: Data analytics platform for processing streams of structured and unstructured data (2014)
5. Frick, T.: *Return on engagement: Content, strategy and design techniques for digital marketing*. Routledge (2013)
6. Garcia-Crespo, A., Colomo-Palacios, R., Gomez-Berbis, J.M., Ruiz-Mezcua, B.: Semo: A framework for customer social networks analysis based on semantics. *Journal of Information Technology* **25**(2), 178–188 (2010)
7. Hoppe, A., Nicolle, C., Roxin, A.: Automatic ontology-based user profile learning from heterogeneous web resources in a big data context. *Proc. VLDB Endow.* **6**(12), 1428–1433 (Aug 2013)
8. Hoppe, A., Roxin, A., Nicolle, C.: Customizing semantic profiling for digital advertising. In: *On the Move to Meaningful Internet Systems: OTM 2014 Workshops*. pp. 469–478. Springer Berlin Heidelberg, Berlin, Heidelberg (2014)
9. Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. *Web Semantics: Science, Services and Agents on the World Wide Web* **23**, 2–15 (2013)
10. Pham, M., Alse, S., Knoblock, C.A., Szekely, P.A.: Semantic labeling: A domain-independent approach. In: *ISWC*. pp. 446–462 (2016)
11. Roman, D., Nikolov, N., Pultier, A., Sukhobok, D., Elvesæter, B., Berre, A., Ye, X., Dimitrov, M., Simov, A., Zarev, M., Moynihan, R., Roberts, B., Berlocher, I., Kim, S., Lee, T., Smith, A., Heath, T.: Datagraft: One-stop-shop for open data management. *Semantic Web* **9**(4), 393–411 (2018)
12. Spahiu, B., Porrini, R., Palmonari, M., Rula, A., Maurino, A.: Abstat: Ontology-driven linked data summaries with pattern minimalization. In: *European Semantic Web Conference*. pp. 381–395. Springer (2016)
13. Sukhobok, D., Nikolov, N., Pultier, A., Ye, X., Berre, A., Moynihan, R., Roberts, B., Elvesæter, B., Mahasivam, N., Roman, D.: Tabular data cleaning and linked data generation with grafterizer. In: *European Semantic Web Conference*. pp. 134–139. Springer (2016)
14. Wertime, K., Fenwick, I.: *DigiMarketing: The essential guide to new media and digital marketing*. John Wiley & Sons (2011)

<sup>22</sup> <https://openweathermap.org>

<sup>23</sup> OWM explicitly recommends to call OWM API by city ID to get unambiguous result for cities. In our pilot we need weather for regions (not available in OWM). In fact, obtaining an ID and hence coordinates from an (ambiguous) toponym is the enrichment problem addressed in our pipeline.