Data enrichment toolchain: a data linking and enrichment platform for heterogeneous data

Luis Sanchez, Jorge Lanza, Juan Ramón Santana, Pablo Sotres, Víctor González, Laura Martín, Gürkan Solmaz, Ernö Kovacs, Maren Dietzel, Anja Summa, et al.

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 August 2023, accepted 17 September 2023, date of publication 20 September 2023,
date of current version 26 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317705

RESEARCH ARTICLE

Data Enrichment Toolchain: A Data Linking and Enrichment Platform for Heterogeneous Data

LUIS SÁNCHEZ[1], JORGE LANZA[1], JUAN RAMÓN SANTANA[1], PABLO SOTRES[1], VÍCTOR GONZÁLEZ[1], LAURA MARTÍN[1], GÜRKAN SOLMAZ[2], (Member, IEEE), ERNÖ KOVACS[2], MAREN DIETZEL[3], ANJA SUMMA[3], AMIR REZA JAFARI[4], ROBERTO MINERVA[4], (Senior Member, IEEE), AND NOEL CRESPI[4], (Member, IEEE)

[1]Network Planning and Mobile Communications Laboratory, Universidad de Cantabria, 39005 Santander, Spain
[2]NEC Laboratories Europe GmbH, 69115 Heidelberg, Germany
[3]Kybeidos GmbH, 69126 Heidelberg, Germany
[4]Samovar, Telecom SudParis, Institut Polytechnique de Paris, 91011 Palaiseau, France

Corresponding author: Luis Sánchez (lsanchez@tlmat.unican.es)

This work was supported in part by the Project Situation-Aware Linked Heterogeneous Enriched Data (SALTED) from the European Union's Connecting Europe Facility Programme under Grant 2020-EU-IA-0274, and in part by the Spanish State Research Agency (AEI) through the Project Semantically-Enabled Interoperable Trustworthy Enriched Data-Spaces (SITED) under Grant PID2021-125725OB-I00.

ABSTRACT Proliferation of data sources associated to Internet of Things (IoT) deployment as well as those bound to Open Data Portals (e.g. European Data Portal, Municipalities Open Data Portals, etc.) and Social Media platforms is creating an abundance of information that is called to bring benefits for both the private and public sectors, through the development of added-value services, increasing administrations' transparency and availability or fostering efficiency of public services. However, pieces of information without a context are significantly less valuable. Raw data lacks semantics and it is highly heterogeneous from one data-source to another. This poses a challenge to make it useful. To turn all this data into valuable information it is necessary to enable its combination so that meaningful context can be created. Moreover, it is fundamental to define the mechanisms enabling the adoption and orchestration of advanced (typically AI-enabled) data processing techniques to be applied over the harmonized datasets and data-streams. This paper presents the Data Enrichment Toolchain (DET) that provides the necessary harmonization and enrichment to datasets and data-streams coming from heterogeneous sources. The value of the enriched data lies on the one hand in the transfer of the data into a semantically grounded knowledge graph and, on the other hand, in the creation of new data through linking, aggregating and reasoning on the data. In both cases, the benefit of employing linked-data modelling and semantics comes from the extension of the metadata that is associated to every piece of information. Furthermore, the experimental evaluation of the DET implementation that we have carried out is also presented in the paper.

INDEX TERMS Data enrichment, semantic annotation, data linking, data processing, heterogenous data, data interoperability.

I. INTRODUCTION

Nowadays, data is becoming the new fuel for economic wealth and creation of novel business models. A multitude of technologies contributes to an abundance of information sources which are already the baseline for many different services. Among them, Internet of Things (IoT), Social Networks and Open Data are probably the most representative ones. However, for an economy of data to flourish there are still several challenges that have to be overcome.

Current solutions are mainly based on centralized approaches where all the information is extracted from the field and forwarded to Data Management Platforms (DMP) in the cloud. Typically, such DMPs are associated with service platforms that allow the development of applications that employ the gathered information. However, existing DMPs

The associate editor coordinating the review of this manuscript and approving it for publication was Rahim Rahmani.

mostly consider the provision of vertical solutions that do not need, and correspondingly are not able, to exploit the abundance of heterogeneous, but related, data. This scenario implies two key problems that form barriers for the adoption of a widely spread economy of data based on level playing field data spaces.

Firstly, the value of data comes from the fact that it allows creating situational awareness which can be further used to take optimized decisions. Moreover, **situational awareness** grows richer as more information sources are used. Thus, there must be solutions in place to guarantee the **interoperability** of the plethora of information sources that can contribute to generate a rich situation awareness that applications can further exploit into optimized decision-making. Moreover, new data sources should be used as they appear. Together, interoperability and **dynamic discoverability** foster competition and innovation as users can seamlessly 'move' to a different provider. Such interoperability should be based on common Application Programming Interfaces (APIs) and data formats, which settles the soft infrastructure of data spaces by providing syntactic interoperability as well as **semantic enrichment and linkage** among pieces of information even if they come from different sources. Moreover, semantic enrichment and linkage should also be exploited to augment each piece of data with important metadata regarding to, among other things, quality, provenance, value, reputation, or limitations of that data.

Secondly, while we have only been able to scratch a bit of the real worth of existing available data, data producers are reluctant to share it. The reason for such situation is that, currently, big corporations are monopolizing the collection of data into centralized platforms for which the rest of players can only accept the terms that these giants fix. Those that are not comfortable with this situation cannot choose a de-centralized solution that lets them fix the access rules to their data, and thus decide to keep the data unavailable, and also unproductive.

In this paper we are presenting the design, implementation and deployment of a Data Enrichment Toolchain (DET) that addresses these challenges and enables harmonization and enrichment of heterogeneous data. The DET leverages the principles of linked data through the adoption of information models and APIs of the Next Generation Service Interfaces Linked Data (NGSI-LD) standard [1] from the European Telecommunication Standards Institute (ETSI). Our goal with the DET is to take existing high value data sources and create new value through the process of ''Data Enrichment''. Data Enrichment can be applied to a multitude of data formats and convert it to a standardized rich data description that make use of the advanced features of NGSI-LD. This approach facilitates data harmonization, and semantic annotation, and the use of Artificial Intelligence (AI) mechanisms to achieve data enrichment.

The advantages that this solution brings forward are that the provided raw data and the enriched data will be more easily accessible to create new services and provide differentiated values that the simple publication of the datasets cannot provide. Additionally, the paper describes the DET implementation that actually enables homogeneous access to highly heterogeneous data sources (e.g. IoT deployments, Open Data portals, Web content or Social Media).

Our main contributions include: (1) the specification of a a modular architecture, based on context management standards, supporting decentralized operation and adaptation to heterogeneous data; (2) the implementation of a toolchain enabling heterogeneous data harmonization and enrichment by leveraging linked-data and AI technologies; and (3) proof of work services validating the proposed platform through actual integration of multiple data sources, and assessing its performance.

Overall, the key innovations that are put forward through the DET that we are describing and evaluating in this paper relates to increasing data value by leveraging semantic annotation to provide some meaning to it, as well as to facilitating discovery of data (specifically, leveraging the NGSI-LD information model), and enabling more complex processing flows and event processing thanks to the establishment of links among pieces of data (existing and/or generated throughout the enrichment process).

The remaining of the paper is structured as follows. Section II presents some related work on data interoperability and semantic enrichment of data. The DET functional architecture description and the specification of its main building blocks are presented in Section III. Following, Section IV presents the details of the DET implementation and deployment that has been done. It also presents the results from the evaluation of functional and non-functional Key Performance Indicators that has been carried out. Finally, Section V concludes the paper.

II. RELATED WORK

In this section we briefly analyze works and initiatives related to two key concepts of the DET value proposition, namely data interoperability and semantic enrichment of data.

A. DATA INTEROPERABILITY

The problem of data interoperability has existed since the early days of information systems. There are several definitions for interoperability in the literature. Among the diverse definitions for interoperability, we quote the ones related to our context. The Cambridge Dictionary gives a general definition for interoperability as ''the ability to work together with other systems or pieces of equipment''. This implies that two interoperable systems can understand one another and use the functionality of each other. In a broader view, interoperability is defined by the Institute of Electrical and Electronics Engineers (IEEE) as ''the ability of two or more systems or components to exchange information and to use the information that has been exchanged'' [2]. According to this definition, interoperability is realized by devising standards. Considering the plethora of data sources that currently exist, in the context of this work, interoperability implies the

ability to transparently access and share the services of such interoperable systems [3].

Currently fragmented data ecosystem is jeopardizing the development of global solutions. The existing multiple parallel platforms have to converge towards offering seamless, global, and linked services to their users. A McKinsey study [4] estimated that a 40% share of the potential economic value of the IoT directly depends on interoperability gaps among IoT platforms. It is necessary to implement solutions that are able to make the already existing data management infrastructures to collaborate in providing a common and portable way of offering their data services. One of the aims of the platform described in this article is to support the automation of the deployment of services/applications over heterogeneous domains.

Interoperability has different facets, embracing architecture, devices, or data. We will focus on the latter one and will show how the proposed DET complements existing works.

In recent years, multiple architectures, frameworks, and layers for interoperability, including semantic approaches, were introduced [5], [6], [7], [8], [9]. They can be catalogued in one of the levels described in the work [10]: connection (basic connectivity and network connectivity), communication (data exchange interoperability) or semantic (understanding in the meaning of the data). Aiming at the definition of an effective DET, we are mainly focusing on those solutions addressing semantic interoperability.

In [11] there is a listing of the key aspects underlying semantic interoperability. More specifically, semantic interoperability applied to IoT is analyzed in [12] and [13] through objects profiling and annotations. A more practical proposition for this kind of interoperability can be found in [14], where interworking proxies are used to accomplish an interoperable behavior between systems based on the NGSI and the oneM2M standards.

Although the works addressing data interoperability share some of our goals related to well-defined data and information models, they have a different approach. While the mentioned proposal focuses on analyzing data commonalities, so that different systems can interact, we are proposing the enforcement of standardized information model and commonly agreed data models capable of leveraging semantic web best practices.

Finally, we have found in the literature some works providing tools to enable interoperability. For example, [15] describes a framework to specify data according to adopted standards. Similarly, a lightweight model-based middleware to simplify interoperability of IoT services is described in [16]. Although the scope of these propositions is different from ours, the formal definition of data model standards could be used by our framework in the future. It is also worth mentioning the development presented in [17], where authors proposed a set of mediation services to access resources in a uniform way in the context of IoT, thus enabling semantic and syntactic interoperability. In this case, the focus is on IoT only and considered the integration of the whole IoT platform, while the DET solution that we are proposing has been designed to enable interoperability and to add the processing steps for data enrichment of datasets and data-streams coming from heterogeneous data sources not limited to IoT platforms. Moreover, DET is focused on the data so existing platforms can be transparently integrated through the corresponding DET injection chain.

B. DATA SEMANTIC ENRICHMENT

The key concept behind the so-called semantic enrichment of data is to augment the value of the data by increasing its contextualized characterization by annotating it with metadata referring to a specific domain of knowledge. An annotation is a form of metadata attached to a dataset or to any of the individual components of it usable to better describe the data type or its information content.

Data enrichment has been widely used in data analysis applications in which the collected data contains limited information and needs to be correlated with existing knowledge to reveal higher-level insights.

In [18], a framework for enriching sensor measurements with semantic concepts is introduced to generate new features. In the Big Active Data project [19], notifications delivered to users can be enriched with other existing data to provide actionable notifications that are individualized per user. Other proposed solutions like [20], focuses on the annotation of data within a collaborative knowledge graph environment, so that metadata is incorporated to data items and annotations remain searchable, and data interconnections are not lost.

A review of other works dealing with semantic enhancement of data should include other pipeline frameworks like the one in Open Semantic ETL toolkit [21] or the SAPP framework [22]. They offer modularized approaches to the creation of data enrichment toolchains that can be extended in a plug-and-play manner.

Some recent similar have proposed analogous architectures for data processing. For example, [23] proposes an IoT data stream processing at edge computing layer, but focuses on investigating its challenges rather than on describing a standard-based solution supporting heterogeneous data (i.e. not only IoT) and dynamically composable enrichment chaining. Interestingly, in this work the performance of the solution is also experimentally evaluated, as we are doing with the DET. While the focus in that work is on the challenges and opportunities of performing data enrichment at the edge, they also focus the evaluation on processing time and comparison of delays as we are doing in Section IV. In [24], so called- Data Acquisition Plans (DAPs), which are dataflows organized according to a direct acyclic graph, are proposed to integrate different information according to domain ontologies on sensors' observations.

Furthermore, recent works have proposed data annotation and enrichment solutions for specific domains, like smart agriculture [25], smart buildings [26], or smart grids [27].

However, we have not found any framework that offers dynamic on-the-fly pipelining, i.e., a set of integrated tools that can help users to enrich the information content of the data independently of its application domain. Thus, dynamically configuring the enrichment process has become a feature on which the DET has specific functionalities.

The goal is providing a platform that users can employ to access semantically enriched data. The DET is an enabler performing enrichment operations. The enrichment of data can be carried out on static data (e.g., existing large data sets) or on a more dynamic context. The DET, in fact, is built to operate on dynamic, continuous, event based, or on-demand data. The enrichment process can be executed on static or runtime data.

III. DATA ENRICHMENT TOOLCHAIN

In this section the DET is described. First the functional architecture is introduced, then the steps in the process of data enrichment are discussed by emphasizing the high-level functions made available by the DET.

A. HIGH-LEVEL DET SPECIFICATION

The DET's main aim is to add value to datasets and data-streams by enriching them through the application of linked-data, semantics, and AI technologies.

1) KEY REQUIREMENTS

The architecture has been defined considering the following premises:

- Data sources can be batch, providing data under request (e.g. RESTful interfaces from an Open Data portal); or real-time, providing data as soon as it is generated (e.g. under publish/subscribe based services). The DET architecture is flexible in order to collect and deal with data. Two operational modes are envisaged:
  - ''batch mode'' accesses already defined data sets and transform them; or
  - ''real-time mode'' collects and process data from sources while they are produced.
- Data must be provided following the Linked Data Design Issues [28].
- Data must be curated, limiting the data garbage provided to applications consuming data from the DET.
- Data can be dynamically linked based on changes in the data (e.g., a newly created entity might result in a *sameAs* link to an existing entity).
- Data can be enriched with new properties as they are created where properties can be new data attributes or new relationships, e.g. to express structural contexts relationships or dynamic situation descriptions.

2) FUNCTIONAL ARCHITECTURE

FIGURE 1 shows the DET functional architecture and presents the flow of data through its building blocks. The DET can be described as the composition of microservices that results in the progressive transformation, formatting, and enhancement of the original information to increase its quality and value. Conceptually, the DET can be understood as a pipeline with a set of components, each one targeting a specific step within a data enrichment process. This chain of transformations is carried out in an iterative manner and some of these steps (and associated software tools and components) can be parametrized in a dynamic way.

The objective of the architecture is to support the needs of different applications that seamlessly consume data collected, stored, prepared, enriched, and managed by the DET. In order to satisfy a general requirement for the DET, the following functions have been identified as key enablers of the architecture: (1) data discovery (i.e., the ability to discover and request the collection of sets and streams of data); (2) data formatting (i.e., the transformation of raw data into well-formed and structured set of data accordingly to data models described in terms of NGSI-LD); (3) data curation (i.e., the identification, and potential correction, of data that do not reflect the expected quality – e.g. outliers, errors in values and the like); (4) data linkage (i.e., the ability to relate different data set according to well-established definitions of relationships); and (5) data enrichment (i.e., the ability to understand and frame the data structures according to situations and contexts and the definition of functions that exploit this contextualization).

The first three of these functions are grouped together in the so-called Injection Chain. Modules belonging to these first three functional categories team up to collect, format, and curate the data that is introduced into the platform. Afterwards, the architecture builds on the NGSI-LD standard, more specifically on its federated and de-centralized context data management API, and enables components from each of the other main functions to add/query or subscribe to data.

Thus, the core component of the DET architecture is the NGSI-LD Context Broker. It enables the linking and enrichment process on top of curated NGSI-LD data, providing access to external applications retrieving and making use of the data. Considering this component as the core of the architecture, security procedures (e.g. OAuth2, JSON Web Tokens, Transport Layer Security, ...) are implemented to ensure that the NGSI-LD interfaces exposed by the Context Broker are safe/secure.

In this sense, it is important to highlight that the decision of employing the NGSI-LD standard as the core of the proposed architecture obeys to the need of addressing the aforementioned key challenges that have to be overcome for the adoption of a widely spread data spaces populated by heterogeneous data sources and where enriched data might be found, namely, in summary, establishing a harmonized manner of representing contextual information enabling the creation of dynamically extendable knowledge graphs.

Thus, employing an NGSI-LD Context Broker as the core component of the DET architecture comes implicit with this design decision. The NGSI-LD Context Broker implements the standardized NGSI-LD protocol to represent and exchange linked data. Through the NGSI-LD API,

context producers and consumers can interact with each other. The NGSI-LD Context Broker, by implementing such API, makes this scenario possible by connecting producers and consumers through the standardized NGSI-LD format and protocol.

The NGSI-LD Context Broker uses the NGSI-LD API and information model to model entities with their properties and relationships, thus forming a property graph with the entities as the nodes. It allows finding information by discovering entities, following relationships and filtering according to properties, relationships, and related meta-information. For data not directly represented in NGSI-LD like video streams or 3D models, links can be added to the model that allows consumers to directly access this information. In this way, Scorpio can provide a graph-based index to a data lake.

The NGSI-LD Context Broker provides several interfaces for querying the stored data so easily analytics can be done on the stored data, like it can be used to predict the situation of an ecosystem. In particular, it enables the following interfaces:

- Create, update, append and delete context information: Providing NGSI-LD Entities for the actual information sharing.
- Query context information, including filtering, geographic scoping, and paging.
- Historical tracking of entity data.
- Subscribe to changes in context information and receive asynchronous notifications.
- Register and discover sources of context information, which allows building distributed and federated deployments.

These interfaces are used in the DET to enable data flow between its components. For example, the entity enrichment component might subscribe to changes in terms of available entities. When a change happens (e.g., a new entity has been created), the entity enrichment component gets notified and receives the newly created entity. It can then perform entity enrichment functions and update the entity with the enriched information as linked metadata to the corresponding entities.

Finally, NGSI-LD Context Brokers can be deployed in a federated setup, either to increase scalability or to achieve some isolation between different organizations. In the DET, this feature has been mainly leveraged to enable separation among the various injection chains so that they can be added or removed from the DET without causing disruption to the DET operation.

B. DATA DISCOVERY AND COLLECTION

Data discovery and collection, as the first stage in the proposed process supported by the architecture, implies the discovery and acquisition of raw data from different data sources. This step is the starting point for the process leading to the curation, certification and distribution of reliable data.

This functionality is used for discovering different types of data from heterogeneous data sources based on application needs. The following is a non-comprehensive list of sources that can be subject of data collection.

- *IoT-based*: data from various IoT sensors in different cities such as traffic, pollution, weather, etc.
- *Social media*: data from famous platforms such as Twitter, Facebook, YouTube, etc.
- *Web-stored*: data from commonly known websites e.g., homepages of companies, businesses or public administration sites.
- *Socioeconomic statistical*: data from open government catalogues to allow for effective public oversight.
- *National and International Meteorological*: data related to weather and climate from different agencies.

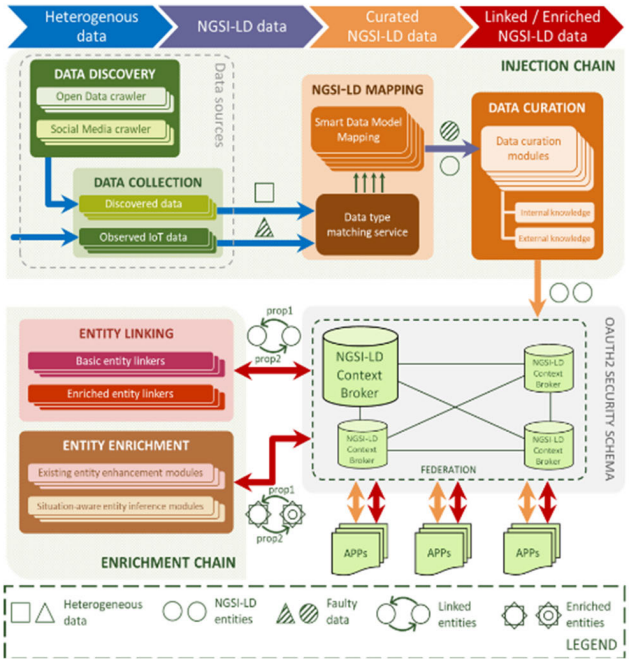

FIGURE 1. DET functional architecture.

Due to the different nature of data sources and the way they generate data, the collection process can be classified as data source dependent. Therefore, different handling mechanisms are developed. For mainly structured, known IoT data, the collection process can be achieved through the implementation of interfaces for asynchronous and synchronous data collection from IoT sensors. Data available from Open Data portals also typically have a formal structure both in terms of data modelling and access interfaces, however, they can have a wide variety from one portal to another even in catalogues referring to the same domains. For social media and web-stored data, the collection process deals with collecting information through data crawlers. They connect to different platforms and collect data by means of official APIs.

The output of this component will be the raw data collected from heterogeneous data sources in different types and formats.

C. DATA MODELLING AND MAPPING

Injection Chains analyze raw data sources and generate the correspondingly normalized data elements using NGSI-LD

information model. By formatting heterogeneous data into a single, standardized format, the mapping function enables uniform processing for the following components in the pipeline.

This component takes the raw data collected in the previous stage as its input, and provides NGSI-LD compliant data as its output. The input, due to its heterogeneous nature, can be represented using several different data formatting standards such as Comma Separated Values (CSV) or JavaScript Object Notation (JSON). Moreover, both the names of the properties and the values can be highly different from one another, as a result of the heterogeneous data sources and their internal policies, language, units, and several other factors.

Three approaches, characterized by increased complexity and offering higher flexibility and scalability, fulfilling this purpose have been identified: (i) The static mapping leverages a pre-established script that maps a specific kind of input data to NGSI-LD. This script requires full domain knowledge of the input data; (ii) The template-based mapping makes use of a set of templates that are filled with the information extracted from the different data sources, generating NGSI-LD compliant output. This approach also requires a priori domain knowledge of the data source and the need for a different template for every output data type, but it enables reusability of the templates and has a more efficient implementation; and (iii) The AI-based mapping uses Machine Learning (ML) techniques to implement the mapping in an automated way. It is split into type identification, template selection, and transformation. In the first step, a trained AI model classifies the input data into multiple categories, corresponding to the potential output data models. Once selected, the corresponding template from the existing pool is used and the transformation is performed.

For example, in the DET implementation that have been carried out (cf. Section IV), one of the implemented mapping modules, the IoT Data Mapper, combines two of the aforementioned approaches, the template-based mapping and AI-based mapping.

It uses a Nearest Neighbor approach to individually map every field of the input document into its corresponding NSGI-LD property in the Smart Data Models. This mapping is based on the words used within that field. Once the NGSI-LD properties are identified, it uses a JMESpath template in order to transform the input document to the pertinent Smart Data Model. These templates work in tandem with a Python class that includes several custom functions to interact with the templates, allowing us to modify the output NGSI-LD document in a more complex manner. For instance, we are able to do type conversion within the document, change the date format, or generate unique identifiers.

However, the data received may not always be known. In order to deal with unknown data, we have implemented an AI-aided type identification module with the Tensorflow and Keras packages for Python. Firstly, we train a model with domain-specific ontologies (in our case, the taxonomy defined in the SmartSantander project) and then this model is used to predict the type of the input heterogeneous data. After the type of input data has been identified, we send it to the Nearest Neighbor mapper along with the newly identified type, which then allows us to select the corresponding JMESpath template as mentioned before. The result is an automatically mapped NGSI-LD entity compliant with its corresponding Smart Data Model.

D. DATA CURATION

The aim of the Data Curation module is to guarantee that the data received, exported by the mappers, is valuable and adequate to be further processed in the DET. Either by tagging the data entities with additional extra information (metadata) or by directly rejecting them in case non-valid data is detected, this module guarantees that linkers, enrichers, and, subsequently, users and applications gathering information from the DET will have not only high-quality data, but a better understanding of its meaning.

The curation process can be considered the first step in the entity enrichment that DET performs. However, it is important to note that within the DET architecture, the Data Curation module is located after the NSGI-LD mapper and just before the data is fully accessible to external applications and further enrichment/linking processes. It acts on the one hand as a kind of firewall that guarantees that only good and coherent data is stored, and on the other ensures that the information to be stored in the Broker has the highest possible informative quality.

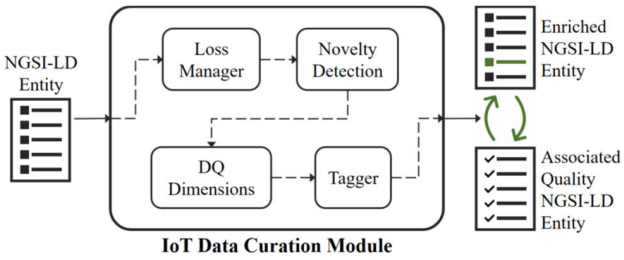

FIGURE 2. Entity linking internal details.

The internal process followed by the curation module is based on four key steps as it is depicted in FIGURE 2. First, it applies completeness verification to detect whether a loss has occurred between two consecutive data items belonging to the same timeseries. If it is found that a loss has occurred, the internal logic of this component will generate a synthetic measurement (NGSI-LD entity) on which the remaining techniques will be performed. Thereafter, outlier detection is performed. Using AI models for short-term prediction, it is determined whether the value of the observation at its timestamp fits or if it can be considered as an anomaly.

To this end, different techniques can be applied for the outlier detection. For example, in the DET implementation carried out, the following algorithms: Isolation Forest (iForest) and Local Outlier Factor (LOF), were used. In the evaluation that was presented in [29], some differences can

be observed between the behavior of the two alternatives. Mainly, the anomalous points detected with the Isolation Forest algorithm correspond to global outliers within the dataset, whereas those detected with the Local Outlier Factor algorithm respond to local anomalies related to the group of neighbors chosen for evaluation.

The last step is dedicated to the acquisition of the Data Quality (DQ) dimensions associated to the incoming data points. Once the DQ features of the NGSI-LD entity being assessed are evaluated and obtained, they are linked to the data entity as a related metadata entity following the DataQualityAssessment data model [30].

E. DATA ENRICHMENT

This subsection describes the data linking and quality/value augmentation. Data linking and enrichment is interchangeably referred to as entity linking and enrichment as well, as the focus in the data models is put on the entities.

1) DATA LINKING

Data linker is defined as "a component that generates NGSI-LD Relationships between two or more NGSI-LD Entities" [31]. The generated Relationship would represent a link between the Entities. Link generation is not restricted in any way but common ways are to discover links are: (1) finding commonalities in the data of the entities or in the data model of entities; and (2) finding spatial associations. The enrichment toolchain might have different data linkers categorized as:

- *IoT data linker* establishes the relation between the entities and their corresponding device of origin. As many data in the toolchain is generated by IoT devices linking data to the devices would enable understanding the infrastructure's status.
- *Geolocation data linker* establishes the relation between the NGSI-LD entities based on their geolocations. For instance, two entities that are located in the same given range or area can be linked to each other.
- *Web data linker* links the entities to the relevant web data results (e.g., crawled data) related to those entities.
- *Correlation linker* establishes the relation between the entities that are correlated based on their data. For instance, intense traffic and pollution can be linked to each other when they are correlated.
- *Semantic linker* establishes the relation between the entities based on their semantic mapping. For instance, two entities that have different names but similar semantic meanings, can be matched.

The number of entities in the NGSI-LD Context Broker is expected to be large. As such, entity matching is facing a scalability problem, as potentially all combinations/pairs of entities need to be evaluated for *sameAs* or other relationships. This complexity is $O(N^2)$, where N is the number of entities in the NGSI-LD Context Broker. To deal with this problem, a three-step pipeline, as shown in FIGURE 3, is leveraged in the DET: (1) A component will filter the entities based on their Smart Data Model type and the available properties in the schema. For the remaining entities, we create all possible combinations; (2) As these might still be too many combinations to perform entity linking in an adequate time-frame, we then block out entity combinations with low likelihood to be matches in a Blocking step; (3) Last, for the remaining candidate entity combinations, we perform a more compute intensive matching/linking step and then push the results back to the NGSI-LD Context Broker.

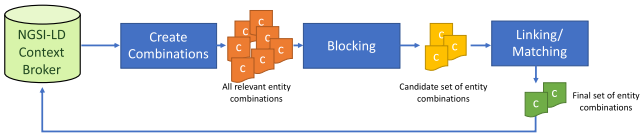

FIGURE 3. Entity linking internal details.

2) DATA QUALITY AND VALUE AUGMENTATION

The enrichment toolchain also considers other ways of data value augmentation. The quality and value augmentation may include additional correlations with other data. For example, web crawling of social media to retrieve relevant sets of data for a specific geo-localized entity can be used to support additional automated analytics such as sentiment analysis. Among the data enrichment modules within the DET, we focus on an example related to the data value augmentation in *smart districts*, through the concept of "*City Liveability Index*".

City Livability Index can assess the status of a district and gives users a way to compare the data from different cities as "indices". The indices are calculated by openly defined methods and accepted practices such as the practices of sustainable development. An index can combine multiple "sub-indices", where each sub-index would be calculated by a metric and the combination would be given by an openly-defined formula. These formulas can be adjusted based on the availability of the data. For instance, the livability in a city can depend on multiple indices such as "RISK, SUSTAINABILITY, or MOBILITY". Each of these indices can take into various factors which would define the sub-indices. For example, MOBILITY may include multiple sub-indices such as PUBLIC_TRANSPORT_USAGE and BIKE_OWNERSHIP.

FIGURE 4 illustrates the concept based on available datasets (FIGURE 4-bottom) such as open datasets or data from European Data Portal (EDP). These datasets can be linked (FIGURE 4-left) through a method such as semantic linking. Data is accessed by different interfaces such as City Livability Index Flexible Frontend (CLIFF) Dashboard or Chatbot (FIGURE 4-top) Similarly, the value-augment data will be provided by the City Livability Indices that are described above. Lastly, the quality of the data can be improved for the districts where the data is limited or suffer from problems such as low resolution. For those scenarios, techniques such as transfer learning can be used to provide accurate predictions (FIGURE 4-right).

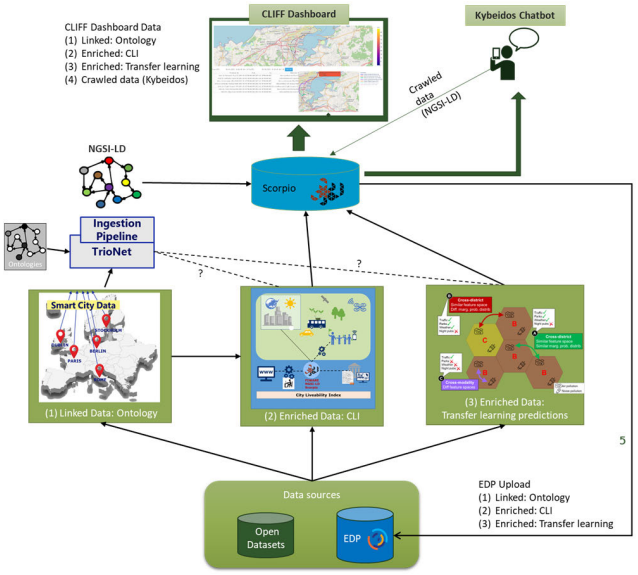

FIGURE 4. View for the concept of the "City Livability Index".

IV. IMPLEMENTATION AND EVALUATION

A. DET DEPLOYMENT DETAILS

The deployment architecture of the DET is based on a federated setup as illustrated in FIGURE 5. This setup includes Scorpio Context Broker Federator that connects the different ''satellites'' of partners in the data plane. In addition, there is a control broker that handles the control plane functions of the DET. Access to the Federator Scorpio Broker and the EMQX Control Broker is restricted using OAuth 2.0. The technology used for the Identity and Access Management is Keycloak.

We have deployed the Scorpio NGSI-LD Brokers from their latest docker images (i.e. scorpiobroker/all-in-one-runner:java-kafka-latest). On the other hand, we have used Python 3 as the programming language for the DET components. These are deployed as separate Python scripts acting independently, which enhances their modularity and reusability. Communications between components are achieved though HTTP, with most components implementing their own lightweight HTTP server with the flask and waitress Python libraries.

Currently, several Injection Chains are implemented: IoT Data Injection Chain, Web Data Injection Chain and Social Media Data Injection Chain. This list can be extended by application developers by introducing new Injection Chains. Technical descriptions and deployment details of the listed injection chains can be found in [31].

B. EVALUATION METHODOLOGY AND KPIS

As part of the deployment that has been carried out for the validation of the DET, several Injection Chains and Data Enrichment modules have been implemented and integrated.

In this section we present the evaluation that has been carried out over some of them. We have taken into consideration two aspects during the execution of several experiments of the deployed DET: (i) the performance of different steps of the DET (specifically, data mapping, data curation and data enrichment); and (ii) the comparison between the distributed approach followed and an alternative centralized counterpart. The performance indicators are, respectively, the processing delay introduced in each DET step, and the relative overhead in terms of extra communication load.

In contrast with the evaluation performed in other studies that also were proposing solutions for semantic interoperability and/or data enrichment, which focused on the qualitative functional assessment [32], [33], [34], we have focused on the quantitative non-functional assessment. In this latter case, as in [23] or [35], the performance assessment, as we have done, is focused on processing time and comparison of delays.

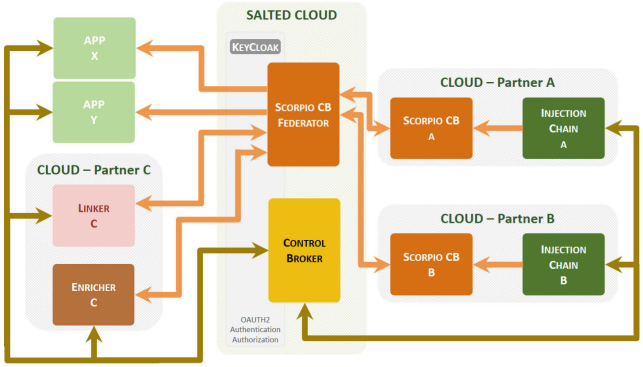

FIGURE 5. Federation setup of DET with satellites.

1) PERFORMANCE EVALUATION OF DET PHASES

The performance indicator analyzed has been, in all cases (i.e. data mapping, curation and enrichment), the processing delay (i.e. the amount of time consumed to process a particular piece of data at each of the DET steps). This delay only has an impact on real-time data (i.e., it does not have major implications for static datasets), and it is comprising the processing time of the single step and the delay to swap data among the DET modules. Moreover, the experiments were not meant to be exhaustive and comprehensively evaluate all the modules that have been implemented. Alternatively, we have evaluated three modules, one of each class.

The Mapper took the measurements from the SmartSantander IoT platform [36]. The Curator assessed several data quality dimensions of each of the measurements in that data stream. Finally, the Linker and the Enricher processed each incoming measurement and added some metadata to them.

The data used for the evaluation consists on temperature measurements generated by the IoT sensors deployed within the SmartSantander framework, in real-time, during several days. These measurements average 230 bytes in length and are formatted in JSON. Over 100 sensors were actively providing measurements every 5 minutes, which, eventually, resulted in roughly one measurement every 3 seconds in average. The content of the measurements comprises the identifier of the originating sensor, its location, the timestamp, the actual temperature value and its unit of measurement.

We have chosen this dataset for several reasons. Firstly, the SmartSantander IoT platform is well-known to us and we are familiar with its reliability. It provides a flow of measurements that matches well with the nature of the evaluation performed. Secondly, temperature measurements are very suitable for an evaluation of the processing time of DET components generally and the data quality metrics specifically. This is mostly due to the statistical nature of the temperature phenomenon itself and the existence of hundreds of sensors within the city of Santander. Lastly, SmartSantander temperature measurements are provided in JSON, which is a well-known format that is widely adopted nowadays in many kinds of data sources.

The experimental set-up consisted on the processing of a large number of measurements (50,000 for the Mapper and the Curator, and 20,000 for the Linker and the Enricher) on an Ubuntu 20.04.5 LTS machine (2 CPU cores, 2.40 GHz clock, 16 GB RAM). The set-up in these cases emulated the deployment of these modules at the same Cloud where the Federator Scorpio NGSI-LD Broker were deployed. As the evaluation is focused on the processing delay, this set-up is meant to avoid introducing further communication delays.

2) CENTRALIZED VS DE-CENTRALIZED DATA STORAGE

As it has been described in the previous section, the deployment carried out follows a distributed approach in which the different Injection Chains are connected to a local instance of the federated NGSI-LD Context Broker. The so-called ''Satellite Brokers''. Among the reasons for adopting such a deployment architecture, the de-centralization of data storage, and the associated upholding of control over their data for the data providers connected to each Injection Chain, as well as the avoidance of large data lakes, and the enablement of edge-based Injection Chains (especially advantageous for IoT data providers), are the key aspects behind this deployment decision.

However, this de-centralization implies an overhead whose assessment has been the focus of the evaluation carried out. Both synchronous (i.e. request-response) and asynchronous (i.e. subscription-notification) queries have been analyzed.

We have focused on the case of IoT sources as they are the ones that are intrinsically decentralized. Thus, there can be actual distribution of both providers and consumers. For the analysis, we have assumed that in a centralized deployment the DET is completely deployed at a general-purpose cloud globally accessible. On contrary, in the decentralized deployment, only the Federator is deployed at a global cloud while the Satellites and their associated Injection Chains are co-located at local Cloudlets close to the IoT infrastructures.

In the centralized case, every time a new observation is generated at its IoT infrastructure, it is sent to the central cloud, while in the decentralized case, the observation is kept at the local Satellite. On consumption, queries are always addressed to the Federator. Synchronous requests are served from the Federator. Directly, in the centralized case, or, indirectly (i.e. after the Federator has gotten it from the corresponding Satellite), in the decentralized one. Alternatively, for asynchronous subscriptions, the Federator forwards the notifications towards the subscribed consumer upon a new observation is generated (again, directly, in the centralized deployment, or, indirectly from the Satellite were the observation is originally stored, in the decentralized case).

The simulation environment included four different providers and four consumers which are located at different Europe regions (i.e. North, South, West and Central Europe). Communication latency across regions is taken from the Inter-Region latency services that major cloud providers like Google or Azure provides. As we are focusing on a relative comparison, absolute latency values are not relevant, but we are normalizing inter-region latencies to: 1 (communication within the same region); 2 (communication between close regions); and 3 (communication between distant regions). The scenario could be extended to more locations just defining more inter-region latencies. Providers generate observations following a Poisson random function. Consumers are subscribed to observations coming from two different providers located in randomly selected regions. The simulations are run 1000 times.

C. RESULTS AND DISCUSSION

Pertaining to the processing overhead introduced by the DET components, Table 1 summarizes the results obtained during the experimental evaluation. The operation of the four modules evaluation has been split in two main phases, Request and Processing. During the Request phase, the module gets all the necessary information to carry out its duties. Once all these data are gathered, during the Processing phase, the actual mapping, curation, linking or enrichment is done. Only in the case of the Mapper, the Request phase is absent, since it only uses the incoming data entity for its operation. As it has been described, for each of the modules a large number of data points (i.e. sensor measurements) were processed. The results presented in Table 1 are the average and standard deviation values for each case.

The overhead introduced is minimal. Only in the case of the Curator, the average processing time raises over 10 milliseconds. Though, for the overall overhead computation, the time that they require to get the necessary information (whether it comes from external sources or from the same Satellite Broker) to perform their operations, which in Table 1 is called request time, has to be considered. While for the Mapper there is no need for additional data, the Curator requires information for feeding the ML-algorithms producing Data Quality metadata and performing observations classification (e.g. outlier or not). Similarly, the evaluated Enricher needs external services to provide the additional attributes with which the data items are augmented. Finally, the Linker also demands information within the Context Broker but, in this case, the request is less demanding.

As from the results shown in Table 1, the example components evaluated show low processing times considering the mid-sized computing resources employed for the

TABLE 1. DET performance evaluation results.

	Mapper	Curator	Linker	Enricher
Avg. processing time (ms)	2.85	15.25	0.16	0.05
Processing time std. dev. (ms)	2.19	0.11	0.02	0.01
Avg. request time (ms)	N/A	306.04	17.40	162.12
Request time std. dev. (ms)	N/A	8.17	10.47	91.98

experimental evaluation. However, for those components that requires larger amounts of data to operate, the main part of the introduced overhead would be attributable to actually getting these data. Thus, looking at these results, it can be concluded that data mapping, which is the only step that has to be mandatorily part of every Injection and Enrichment chain within the DET has a negligible overhead. Moreover, data linking, which is the minimum form of data augmentation, can also be always enabled due to its very low impact. On contrary, the inclusion of data curation or complex enrichment of data items might need to be carefully assessed considering the functionality-vs-overhead trade-off. In these cases, dynamic composition of the DET would allow consumers to choose between higher performance or enriched data.

Regarding the overhead comparison between the centralized and the de-centralized deployments, Table 2 presents the results of the evaluation carried out. Since the objective of the evaluation was to compare the two alternatives, the most usual centralized approach where data is gathered at one server located in the cloud, and the de-centralized one chosen for the DET design and implementation, the values shown in Table 2 have been normalized to the average inter-region latency, so that they are not dependant on the actual evaluation scenario but they can be extrapolated. As the evaluation use a Monte Carlo strategy with 1000 repetitions of the simulations, Table 2 presents the average and standard deviation values of the normalized delay for the case of both synchronous (Synch) requests (i.e. the time taken between a data request by a consumer and its corresponding response) and asynchronous (Asynch) notifications (i.e. the time taken between a data publication by a provider and its corresponding notification to the subscribed consumer).

The centralized approach outperforms the de-centralized one, mainly for the synchronous requests case. However, for the asynchronous subscriptions the advantage of the centralized approach is smaller. This is mainly because, in this case, in which the publication of data items is decoupled from the notification to the interested subscribers, the de-centralized option benefits from the low delay introduced in the publication process, which is always handled locally, and the fact that those data items that do not have a matching subscription are not sent to the central broker, thus introducing valuable savings both in terms of delay and in terms of data transmission overhead.

TABLE 2. Centralized-vs-decentralized overhead comparison.

	Centr. Synch.	Decentr. Synch.	Centr. Asynch.	Decentr. Asynch.
Avg. delay (normalized)	3.5	7	2.63	3
Delay std. dev. (normalized)	0.0038	0.0054	0.0045	0.0085

V. CONCLUSION

The DET has been designed to enable heterogeneous data harmonization and enrichment. It leverages linked-data and AI-based data processing. Its modular architecture, for which the NGSI-LD context management standard is the key baseline, is meant to support decentralized operation and adaptation to heterogeneous data. As it has been described in the paper, the proposed solution is able to extract high value datasets from existing heterogeneous data sources (from public and private stakeholders) and publish enriched datasets and data-streams using NGSI-LD, which have been harmonized and aggregated using a data enrichment toolchain.

Overall, we argue that the semantic data enrichment that is supported by the DET brings many advantages to data processing, such as: (1) making data more valuable by providing some meaning to it; (2) enabling more efficient and accurate discovery and reasoning; (3) establishing links among pieces of data; and, (4) allowing more complex processing flows and event processing.

The proposed solution has been validated through actual integration of multiple data sources. The actual DET implementation and deployment has been also briefly presented and it has been the basis for the experimental evaluation that has been performed for assessing its performance. The evaluation has shown that the de-centralized approach adopted for the DET deployment enables larger control for data providers while it does not jeopardize the performance. In this respect, the design decision taken related to the decentralized deployment of the DET has also been proven as appropriate, as the evaluation results have demonstrated that the overhead introduced compared to more classical centralized data management platforms is not that high considering the non-functional benefits of keeping data within the provider's domain (e.g. sovereignty of data, exploitation of edge infrastructures, etc.). Moreover, the results from this evaluation have demonstrated the viability of the proposed solution not only from the functional standpoint but also from the non-functional one, as the delay introduced by the data enrichment modules implemented can be considered negligible.

REFERENCES

[1] *Context Information Management (CIM) European Telecommunications Standard Institute, Industry Specification Group (ISG)*, NGSI-LD API, Sophia Antipolis, France, 2021.

[2] J. Radatz, A. Geraci, and F. Katki, *IEEE Standard Glossary of Software Engineering Terminology*, Computer Society of the IEEE, Standard IEEE Std. 610.12-1990, 1990.
[3] J. Kiljander, A. D'elia, F. Morandi, P. Hyttinen, J. Takalo-Mattila, A. Ylisaukko-Oja, J.-P. Soininen, and T. S. Cinotti, "Semantic interoperability architecture for pervasive computing and Internet of Things," *IEEE Access*, vol. 2, pp. 856–873, 2014, doi: 10.1109/access.2014.2347992.
[4] *The Internet of Things: Mapping the Value Beyond the Hype*, McKinsey Global Inst., New York, NY, USA, 2015.
[5] A. Bröring, S. Schmid, C.-K. Schindhelm, A. Khelil, S. Käbisch, D. Kramer, D. Le Phuoc, J. Mitic, D. Anicic, and E. Teniente, "Enabling IoT ecosystems through platform interoperability," *IEEE Softw.*, vol. 34, no. 1, pp. 54–61, Jan. 2017, doi: 10.1109/ms.2017.2.
[6] P. Glanon, S. Azaiez, and C. Mraidha, "A modular interoperability layer for connecting the business and manufacturing systems," in *Proc. 14th IEEE Int. Workshop Factory Commun. Syst. (WFCS)*, Jun. 2018, pp. 1–4, doi: 10.1109/wfcs.2018.8402383.
[7] O. Borgogno and G. Colangelo, "Data sharing and interoperability: Fostering innovation and competition through APIs," *Comput. Law Secur. Rev.*, vol. 35, no. 5, Oct. 2019, Art. no. 105314, doi: 10.1016/j.clsr.2019.03.008.
[8] J. Nilsson and F. Sandin, "Semantic interoperability in industry 4.0: Survey of recent developments and outlook," in *Proc. IEEE 16th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2018, pp. 127–132, doi: 10.1109/indin.2018.8471971.
[9] S. Bader, "The international data spaces information model—An ontology for sovereign exchange of digital content," in *Proc. Int. Semantic Web Conf.*, in Lecture Notes in Computer Science, 2020, pp. 176–192, doi: 10.1007/978-3-030-62466-8_12.
[10] S. Pantsar-Syväniemi, A. Purhonen, E. Ovaska, J. Kuusijärvi, and A. Evesti, "Situation-based and self-adaptive applications for the smart environment," *J. Ambient Intell. Smart Environments*, vol. 4, no. 6, pp. 491–516, 2012, doi: 10.3233/ais-2012-0179.
[11] S. D. Nagowah, H. B. Sta, and B. A. Gobin-Rahimbux, "Towards achieving semantic interoperability in an IoT-enabled smart campus," in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Oct. 2019, pp. 593–598, doi: 10.1109/ISC246665.2019.9071694.
[12] A. Mazayev, J. A. Martins, and N. Correia, "Interoperability in IoT through the semantic profiling of objects," *IEEE Access*, vol. 6, pp. 19379–19385, 2018, doi: 10.1109/access.2017.2763425.
[13] N. Villanueva-Rosales, L. Garnica-Chavira, V. M. Larios, L. Gómez, and E. Aceves, "Semantic-enhanced living labs for better interoperability of smart cities solutions," in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Sep. 2016, pp. 1–2, doi: 10.1109/isc2.2016.7580775.
[14] J. An, F. Le Gall, J. Kim, J. Yun, J. Hwang, M. Bauer, M. Zhao, and J. Song, "Toward global IoT-enabled smart cities interworking using adaptive semantic adapter," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5753–5765, Jun. 2019, doi: 10.1109/jiot.2019.2905275.
[15] J. L. Hernandez, R. García, M. Fischer, J. Schonowski, D. Atlan, and T. Ruohomäki, "An interoperable open specifications framework for smart city urban platforms," in *Proc. Global IoT Summit (GIoTS)*, 2019, pp. 1–7, doi: 10.1109/giots.2019.8766396.
[16] P. Grace, J. Barbosa, B. Pickering, and M. Surridge, "Taming the interoperability challenges of complex IoT systems," in *Proc. 1st ACM Workshop Middleware Context-Aware Appl. IoT*, Dec. 2014, pp. 1–9, doi: 10.1145/2676743.2676744.
[17] I. P. Zarko, S. Mueller, M. Plociennik, T. Rajtar, M. Jacoby, and M. Pardi, "The symbiote solution for semantic and syntactic interoperability of cloud-based IoT platforms," in *Proc. Global IoT Summit (GIoTS)*, 2019, pp. 1–6, doi: 10.1109/giots.2019.8766420.
[18] A. Moraru and D. Mladenic, "A framework for semantic enrichment of sensor data," *J. Comput. Inf. Technol.*, vol. 20, no. 3, 2012, Art. no. 1002093, doi: 10.2498/cit.1002093.
[19] S. Jacobs, M. Y. S. Uddin, M. Carey, V. Hristidis, V. J. Tsotras, N. Venkatasubramanian, Y. Wu, S. Safir, P. Kaul, X. Wang, M. A. Qader, and Y. Li, "A bad demonstration: Towards big active data," *Proc. VLDB Endowment*, vol. 10, no. 12, pp. 1941–1944, 2017, doi: 10.14778/3137765.3137814.
[20] L. Rossenova, Z. Schubert, R. Vock, L. Sohmen, L. Günther, P. Duchesne, and I. Blümel, "Collaborative annotation and semantic enrichment of 3D media," in *Proc. 22nd ACM/IEEE Joint Conf. Digit. Libraries*, Jun. 2022, pp. 1–5, doi: 10.1145/3529372.3533289.
[21] *Open Semantic Search, Open Semantic ETL Toolkit for Data Integration, Data Analysis, Document Analysis, Information Extraction and Data Enrichment*. Accessed: Aug. 20, 2023. [Online]. Available: https://www.opensemanticsearch.org/etl
[22] J. J. Koehorst, J. C. J. van Dam, E. Saccenti, V. A. P. Martins dos Santos, M. Suarez-Diez, and P. J. Schaap, "SAPP: Functional genome annotation and analysis through a semantic framework using FAIR principles," *Bioinformatics*, vol. 34, no. 8, pp. 1401–1403, Apr. 2018, doi: 10.1093/bioinformatics/btx767.
[23] F. Xhafa, B. Kilic, and P. Krause, "Evaluation of IoT stream processing at edge computing layer for semantic data enrichment," *Future Gener. Comput. Syst.*, vol. 105, pp. 730–736, Apr. 2020, doi: 10.1016/j.future.2019.12.031.
[24] S. Valtolina, L. Ferrari, and M. Mesiti, "Ontology-based consistent specification of sensor data acquisition plans in cross-domain IoT platforms," *IEEE Access*, vol. 7, pp. 176141–176169, 2019, doi: 10.1109/access.2019.2957855.
[25] P. Mylonas, Y. Voutos, and A. Sofou, "A collaborative pilot platform for data annotation and enrichment in viticulture," *Information*, vol. 10, no. 4, p. 149, Apr. 2019, doi: 10.3390/info10040149.
[26] T. Wang, V. J. L. Gan, D. Hu, and H. Liu, "Digital twin-enabled built environment sensing and monitoring through semantic enrichment of BIM with SensorML," *Autom. Construction*, vol. 144, Dec. 2022, Art. no. 104625, doi: 10.1016/j.autcon.2022.104625.
[27] A. Fernández-Izquierdo, A. Cimmino, C. Patsonakis, A. C. Tsolakis, R. García-Castro, D. Ioannidis, and D. Tzovaras, "OpenADR ontology: Semantic enrichment of demand response strategies in smart grids," in *Proc. Int. Conf. Smart Energy Syst. Technol. (SEST)*, Sep. 2020, pp. 1–6, doi: 10.1109/sest48500.2020.9203093.
[28] T. Berners-Lee. (Jul. 27, 2006). *Linked Data—Design Issues*. [Online]. Available: https://www.w3.org/DesignIssues/LinkedData
[29] L. Martín, L. Sánchez, J. Lanza, and P. Sotres, "Development and evaluation of artificial intelligence techniques for IoT data quality assessment and curation," *Internet Things*, vol. 22, Jul. 2023, Art. no. 100779, doi: 10.1016/j.iot.2023.100779.
[30] GitHub Repository. *Data Quality Assessment, Smart Data Model*. Accessed: Aug. 20, 2023. [Online]. Available: https://github.com/smart-data-models/dataModel.DataQuality
[31] V. González, L. Martín, J. Lanza, and J. R. Santana, 'Situation-aware linked heterogeneous enriched data (SALTED): D2.2: Report on data modelling and linking," Project Deliverable, European Union, Brussels, Luxembourg, Tech. Rep., 2020-EU-IA-0274, 2023.
[32] S. Aydin and M. N. Aydin, "Semantic and syntactic interoperability for agricultural open-data platforms in the context of IoT using crop-specific trait ontologies," *Appl. Sci.*, vol. 10, no. 13, p. 4460, Jun. 2020, doi: 10.3390/app10134460.
[33] P. Fafalios, K. Petrakis, G. Samaritakis, K. Doerr, A. Kritsotaki, Y. Tzitzikas, and M. Doerr, "FAST CAT: Collaborative data entry and curation for semantic interoperability in digital humanities," *J. Comput. Cultural Heritage*, vol. 14, no. 4, pp. 1–20, Dec. 2021, doi: 10.1145/3461460.
[34] P. H. L. Rettore, B. P. Santos, R. Rigolin F. Lopes, G. Maia, L. A. Villas, and A. A. F. Loureiro, "Road data enrichment framework based on heterogeneous data fusion for ITS," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1751–1766, Apr. 2020, doi: 10.1109/TITS.2020.2971111.
[35] N. Kalatzis, G. Routis, Y. Marinellis, M. Avgeris, I. Roussaki, S. Papavassiliou, and M. Anagnostou, "Semantic interoperability for IoT platforms in support of decision making: An experiment on early wildfire detection," *Sensors*, vol. 19, no. 3, p. 528, Jan. 2019, doi: 10.3390/s19030528.
[36] L. Sanchez, L. Muñoz, J. A. Galache, P. Sotres, J. R. Santana, V. Gutierrez, R. Ramdhany, A. Gluhak, S. Krco, E. Theodoridis, and D. Pfisterer, "SmartSantander: IoT experimentation over a smart city testbed," *Comput. Netw.*, vol. 61, pp. 217–238, Mar. 2014, doi: 10.1016/j.bjp.2013.12.020.

LUIS SÁNCHEZ received the M.Sc. and Ph.D. degrees in telecommunications engineering, in 2002 and 2009, respectively. He is currently an Associate Professor with Universidad de Cantabria, Spain. He is also active on the IoT-enabled smart cities and the application of AI for data enrichment. He has led and/or participated in more than 15 projects belonging to different EU framework programs. He has authored more than 60 papers at international journals and conferences. He often participates in panels discussing about innovation supported by IoT in smart cities. He is also an expert of several European countries national funding agencies.

JORGE LANZA received the Ph.D. degree in telecommunications engineering from the University of Cantabria (UC), Spain, in 2014. He is currently an Associate Professor with the Network Planning and Mobile Communications Laboratory, UC. He has participated in several research projects, national and international, with both private and public funding. His research interest includes the IoT infrastructures toward federating deployments in different locations using semantics technologies. In addition, his work has included combined mobility and security for the wireless internet.

JUAN RAMÓN SANTANA received the Ph.D. degree in telecommunication engineering from the University of Cantabria (UC), in 2021. He has participated in a number of international projects, mostly within the Internet of Things (IoT) paradigm and its application in the smart city domain. He is currently a Research Fellow with the Network Planning and Mobile Communications Laboratory, UC. He has authored more than 30 publications, including conferences, journals, and book chapters. His research interests include the data management plane of IoT infrastructures and their federation using semantic-enabled technologies.

PABLO SOTRES received the Ph.D. degree in telecommunications engineering from the University of Cantabria (UC), Spain, in 2008 and 2021, respectively. He is currently a Senior Research Fellow with the Network Planning and Mobile Communications Laboratory, which belongs to the Communications Engineering Department, UC. He has been involved in several different international projects framed under the smart city paradigm, such as SmartSantander and related to inter-testbed federation, such as Fed4FIRE, Fed4FIRE+, and Wise-IoT.

VÍCTOR GONZÁLEZ received the master's degree in telecommunications engineering from Universidad de Cantabria (UC), Spain, in 2021. He is currently a Research Fellow with the Network Planning and Mobile Communications Laboratory, UC. His research interests include the application of blockchain technology for the distributed and secure sharing of IoT data. Moreover, he is also active on applying semantic web principles to data sharing and, this way, developing a fully distributed secure data sharing ecosystem.

LAURA MARTÍN received the master's degree in telecommunications engineering from Universidad de Cantabria (UC), Spain, in 2022, where she is currently pursuing the Ph.D. degree in telecommunications engineering. She is also a Research Fellow with the Network Planning and Mobile Communications Laboratory, UC. Her research interest includes the application of artificial intelligence for the enrichment of IoT data. Moreover, she is also active on applying semantic web principles to data sharing and, this way, developing a fully distributed enriched data sharing ecosystem.

GÜRKAN SOLMAZ (Member, IEEE) received the Ph.D. degree from the University of Central Florida. He is currently a Senior Researcher with the NEC Laboratories Europe, Germany, where he applies to skills toward mobile computing, AI/ML, and cloud-edge systems for IoT applications in smart cities and smart mobility. His research interests include wireless ad-hoc and sensor networks and human mobility models. He is a member of the ACM Future of Computing Academy.

ERNÖ KOVACS received the Ph.D. degree from the University of Stuttgart. He is currently the Senior Manager of the NEC Laboratories Europe, Heidelberg, Germany. He is also leading the research group "Cloud Services and Smart Things." His team is involved in oneM2M standardization, the European Future Internet Core Platform FIWARE, and many international research projects. He holds 27 granted patents. His research interests include context brokering, cloud-edge computing, real-time situation awareness, knowledge extraction, and smart cities.

MAREN DIETZEL received the Diploma degree in industrial engineering from the Dresden University of Technology, in 2021. She is currently an IT Consultant with Kybeidos GmbH, specializing in the development of PoC architectures and applications in the field of data science. Her research interests include making linked data understandable to the public and advancing innovative and intelligent applications that leverage this.

ANJA SUMMA received the master's degree in computational linguistics from the University of Heidelberg, in 2016. She is currently an IT Consultant with Kybeidos GmbH, specializing in data science communication and NLP. Her research interests include creating innovative applications that benefit local citizens, but can be used all across Europe and making linked data understandable to the public.

AMIR REZA JAFARI received the B.Sc. degree in computer engineering from the Isfahan University of Technology, Iran, in 2018, and the M.Sc. degree in information technology—multimedia systems from Tehran University, Iran, in 2021. He is currently pursuing the Ph.D. degree with the Data Intelligence and Communication Engineering Laboratory (DICE), Telecom SudParis, Institut Polytechnique de Paris, France. His research interests include social network analysis, data science, emotion detection, and knowledge graphs.

ROBERTO MINERVA (Senior Member, IEEE) received the Ph.D. degree in computer science and telecommunications from UPMC (Paris-Sorbonne University), Paris, France, in 2013. He is currently an Associate Professor with the Service Architecture Laboratory, Institut Mines Telecom—Telecom Sud Paris, Institute Polytechnique de Paris, Paris. From 1987 to 2016, he was a Researcher and then a responsible of the service architectures and network intelligence area with the Telecom Italia Research Center. From 2016 to 2018, he was the Technical Project Leader of SoftFIRE, a European Project devoted to the experimentation of NFV, SDN, and edge computing. He has been the Chairperson of the IEEE IoT Initiative (2014–2016).

NOEL CRESPI (Member, IEEE) received the master's degree from the Universities of Orsay and Canterbury, the Diplome d'ingénieur degree from Telecom ParisTech, and the Ph.D. and Habilitation degrees from Paris VI University. He joined the Institut Mines-Telecom, in 2002. He is currently a Professor and the M.Sc. Program Director of Institut Polytechnique de Paris, Telecom SudParis, where he is leading the Service Architecture Laboratory. He coordinates the standardization activities with the Institute Telecom at ETSI, 3GPP, and ITU-T. He is also an Adjunct Professor with KAIST, South Korea, an affiliate Professor with Concordia University, Canada, and a Guest Researcher with the University of Goettingen, Germany. He is also the Scientific Director of the French-Korean Laboratory ILLUMINE. He is the author/coauthor of 400 articles and contributions in standardization. His current research interests include data analytics, the Internet of Things, and softwarization.