Towards Big Data Analytics in Large-Scale Federations of Semantically Heterogeneous IoT Platforms

Ilias Kalamaras$^{(\boxtimes)}$, Nikolaos Kaklanis, Kostantinos Votis,
and Dimitrios Tzovaras

Information Technologies Institute, Centre for Research and Technology Hellas,
Thessaloniki, Greece
{kalamar,nkak,kvotis,tzovaras}@iti.gr

Abstract. The technological advances in the Internet-of-Things (IoT) have led to the generation of large amounts of data and the production of a large number of IoT platforms for their management. The abundance of raw data necessitates the use of data analytics in order to extract useful patterns for decision making. Current architectures for big data analytics in the IoT domain address the large volume and velocity of the produced data. However, they do not address the semantic heterogeneity in the data models used by diverse IoT platforms, which emerges when large-scale deployments, spanning across multiple deployment sites, are considered. This paper proposes an architecture for big data analytics in the context of large-scale IoT systems consisting of multiple IoT platforms. A Semantic Interoperability Layer (SIL) handles the interoperability among the data models of the individual platforms, using semantic mappings between them and a unified ontology. Data queries to the SIL and result collection is handled by a cloud-based data management layer, namely the Data Lake, along with storage of metadata needed by data analytics methods. Based on this infrastructure, web-based data analytics and visual analytics methods are used to analyze the collected data, while being agnostic of platform-specific details. The proposed architecture is developed in the context of healthcare provision for older people, although it can be applied to any IoT domain.

Keywords: Internet-of-Things · Big data analytics
Semantic interoperability · Healthy ageing

1 Introduction

The technological advances of the Internet-of-Things (IoT) have led to the development of human-centric IoT applications, such as e-Health and intelligent transportation systems. Such applications allow the collection of valuable domain information, assisting human operators and decision makers in providing services

for the well-being of individuals and communities. In the context of e-Health, for example, IoT systems can be used to collect data from a large number of patients, allowing a close monitoring of their health and the provision of (automatic or not) interventions, or the formation of health management policies. Currently running European research projects, such as ACTIVAGE [1] and FrailSafe [2], in an attempt to address the healthcare needs of the increasing number of ageing population, provide promising solutions towards the use of IoT technologies for older people monitoring and assistance.

The extensive use of IoT technologies has led to two outcomes. First, there is a very large volume of data (collected by a wide variety of sensors, such as wearables, environmental sensors, appliance usage monitors, etc.), which exceeds the storage and processing power limits of stand-alone applications (big data). Second, there is a growing number of developed IoT platforms providing off-the-self solutions for the development and deployment of IoT applications, without the need for extensive programming. However, in large-scale applications, spanning a large number of different installations, maybe across different countries, each installation may use a different IoT platform, having its own data model for describing the IoT devices and collected data. These models are often incompatible with each other in terms of semantics, making the necessity of semantic interoperability apparent. There is a need to have a common semantic model for describing the concepts of all IoT platforms, in order for large-scale data analytics methods to perform.

This paper proposes an architecture that allows big data analytics methods to perform on large-scale IoT deployments, spanning multiple diverse IoT platforms. Interoperability among the IoT platforms is handled by the introduced *Semantic Interoperability Layer (SIL)*, providing a unified data model and semantic mappings. Data analytics methods are supported by the introduced *Data Lake*, which is based on the SIL and maintains its own cloud storage for extracted features, trained models and any metadata that are needed by analytics methods. The architecture is developed in the context of the ACTIVAGE project [1], whose goal is to support large-scale IoT applications in deployment sites across European countries, in order to exploit the large volume of collected data. Towards this goal, existing IoT platforms already deployed in different sites are used, as well as various sensing systems, such as the behavioural monitoring systems developed in the FrailSafe project [2]. Providing an infrastructure for combining the diverse platforms and data models can provide large-scale data analytics for assisting older people, clinicians and researchers.

The rest of this paper is organized as follows. Section 2 presents background, work regarding big data analytics and semantic interoperability. Section 3 describes the proposed architecture, covering the Semantic Interoperability Layer, the Data Lake, and the data analytics and visualization components. Section 4 describes scenarios for the preliminary evaluation of the proposed architecture, while Sect. 5 concludes the paper, providing information about the next steps.

2   Background

2.1   Big Data Analytics

Data analytics aim at analyzing raw data, in order to extract information that is more meaningful and valuable to the human operator in order to understand the data and make decisions. In the context of IoT, data analytics are mostly concerned with classification, clustering and high-level data representation [3]. *Classification* methods assign an observation to one of multiple classes, after being trained using data with known classes. Common classification methods currently used include Support Vector Machines (SVM) [4], and Random Forests [5]. *Anomaly detection* methods detect unusual circumstances by classifying observations as normal or abnormal, e.g. Local Outlier Factor (LOF) [6] and Bayesian Robust PCA (BRPCA) [7]. *Clustering* methods split observations in groups of similar characteristics, without using training information [8]. Hierarchical clustering proceeds by recursively joining or separating observations, until a tree-like structure is formed, while partitioning clustering, such as k-means and k-medoids, considers an arbitrary starting split, iteratively updating it to best represent the data. Methods to construct *high-level representations* for raw data can remove unnecessary or redundant dimensions. Principal Component Analysis (PCA) [9], Multi-Dimensional Scaling (MDS) [10] and graph embedding methods [11], attempt to find subspaces (manifolds) of maximum information and minimum dimension inside the raw data space. In the context of time series analysis, ARMA models [12] and variants are used to extract high-level information, such as trends and periodicities, from the raw data.

Several architectures for big data analytics in IoT applications have been proposed. The authors of [3] provide a related review and propose an architecture where the data collected by sensors are stored in cloud databases, allowing large-scale data analytics methods to operate on them, using cluster computing frameworks, such as Apache Spark [13] and Hadoop [14]. The authors of [15] propose a framework for off-line and on-line analysis of IoT data of large volume and velocity, by computing model parameters off-line and using them for real-time analysis. The authors of [16] propose a 4-tier architecture, covering data generation by sensors, communication between sensors and gateways, data analysis using cluster computing, and finally data interpretation by human operators. Most big data analytics architectures in the IoT domain are concerned with handling the large volume and velocity of the produced data, without addressing the variety and heterogeneity in their semantics.

2.2   Semantic Interoperability

There is a large number of IoT platforms for managing devices and data, each using a different ontology to describe its semantics [17]. The SSN (Semantic Sensor Network) ontology [18] describes sensors in terms of their functionalities, measurements and deployments, although it has limitations regarding real-time

data collection. The oneM2M ontology [19] has been supported by IoT standardization bodies, although it also has limitations in terms of contextual data annotation. The IoTivity platform [20] is based on the models of the Open Connectivity Foundation (OCF) [21], which aims at providing a common framework for communication among IoT devices and gateways. The OpenIoT ontology [22], utilized by the OPENIoT platform [23], is based on the SSN ontology and adds concepts related to IoT applications and testbeds. The IoT-Lite ontology [24], used by the FIWARE platform [25], is a recent attempt to collect existing concepts of the IoT domain in a common ontology. Ad-hoc data models have also been built for the purposes of various existing open-source IoT platforms, including sensiNact [26], universAAL [27], Sofia2 [28] and SENIORSome [29].

This abundance of IoT ontologies creates interoperability issues in large-scale applications, where IoT platforms with different ontologies must cooperate. *Semantic interoperability* ensures that all components have a common understanding of the meaning of the information being exchanged [30]. Attempts have been made to promote semantic interoperability by unifying existing ontologies. The SAREF (Smart Appliance REFerence) ontology [31] is such an attempt, unifying concepts from several ontologies in the smart appliances domain, in order to cover larger applications. The authors of [32] use the ontology interconnection methodology of [33], in order to unify existing ontologies in the IoT domain, within the context of the FIESTA-IoT European project [34].

The above review suggests that architectures for big data analytics in IoT systems do exist, but they focus on handling the large data volume and velocity, without addressing the heterogeneity of the available data models. Attempts to address heterogeneity are being made, but they are not targeted to providing a basis for large-scale data analytics methods. The current paper aims to contribute to this direction, by proposing an architecture for large-scale IoT data analytics, based on semantic interoperability across diverse IoT platforms.

3   The ACTIVAGE Data Analytics Architecture

The proposed architecture for large-scale data analytics is depicted in Fig. 1b. It is based on the structure of existing IoT frameworks, as depicted in Fig. 1a, forming a stack of layers ranging from the IoT devices at the bottom, towards data analytics and visualization at the top. However, instead of a single IoT platform to handle the devices at the bottom, there are now many platforms, each operating separately, with its own devices, data storage and component semantics. The following IoT platforms are considered in ACTIVAGE, although any number of platforms is supported: FIWARE [25], sensiNact [26], universAAL [27], IoTivity [20], Sofia2 [28], SENIORSome [29] and OPENIoT [23]. The next layer is the *Semantic Interoperability Layer (SIL)*, which unifies the ontologies of the IoT platforms and offers common semantics for their components. The presence of the SIL eliminates any issues of compatibility between inter-platform hardware and software, as each platform manages its own hardware and software, in order to collect data. Interoperability in ACTIVAGE happens in a conceptual
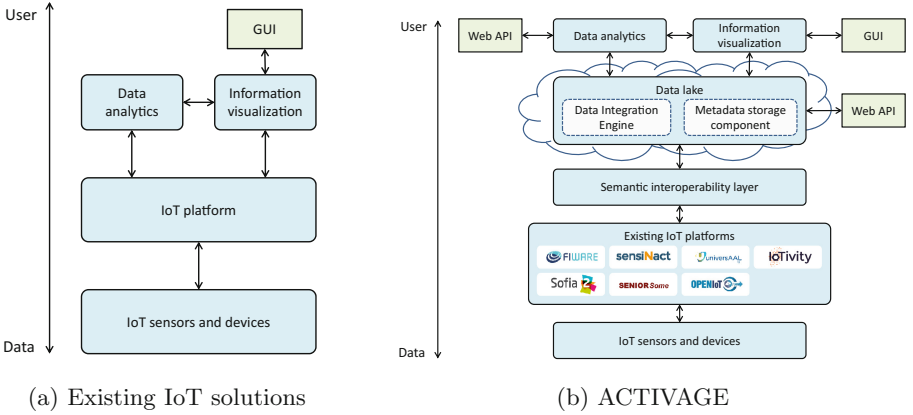
(a) Existing IoT solutions

(b) ACTIVAGE

Fig. 1. The proposed ACTIVAGE architecture (b) extends existing IoT solutions (a), by adding layers regarding platform interoperability and data management.

level, by ensuring the compatibility between different data representations, using the SIL semantic mappings. Above the SIL is the *Data Lake*, which, through its Data Integration Engine, directs the queries coming from the upper layers towards the SIL and collects the data retrieved from the IoT platforms. The Data Lake also contains a Metadata Storage component, for storing metadata (models, etc.) produced and needed by the data analytics methods. The Data Lake components are cloud-based, offering Web APIs for their usage. Based on the infrastructure of the SIL and the Data Lake, the top layers, data analytics and information visualization, can operate, extracting patterns and producing visualizations through Web APIs and graphical interfaces.

3.1 Semantic Interoperability Layer

The Semantic Interoperability Layer (SIL) is responsible for providing an abstraction for the representation of devices, attributes and data, that is agnostic of any IoT platform-specific details and naming conventions. In order to provide interoperability, the SIL maintains a common ontology describing the components of an IoT platform, namely the *ACTIVAGE ontology*. This ontology unifies the ontologies of the participating IoT platforms, so that common names are given for concepts with the same semantics. Platform-specific data representations may be both structured (schema-based databases), or unstructured (schema-less databases). The SIL provides semantic mappings between the common unified model and these individual data models of the IoT platforms.

The ACTIVAGE ontology is based on existing IoT ontologies, such as SSN [18], SAREF [31], oneM2M [19], IoT-Lite [24] and OpenIoT [22], and aims to combine and extend them. It defines basic concepts of IoT platforms, such as *Device* (a physical object able to communicate with its environment), *Service* (a software component able to perform some functionality) and *Measurement* (a

piece of information collected by a device). Some concepts, such as "Device", are widely used across many existing IoT ontologies, while others, such as "Service" and "Measurement", are defined only in some of them. The ACTIVAGE ontology aims at gathering both widely used and less used concepts, in order to cover the types of applications built on top of ACTIVAGE, such as data analytics. The ACTIVAGE ontology is currently under development and is meant to be constantly developed as the proposed architecture is evaluated in real-world scenarios and further IoT platforms are integrated.

3.2   Data Lake

The Data Lake acts as an intermediate layer between the Semantic Interoperability Layer and the data analytics and visualization methods above. It consists of the following components:

- The **Data Integration Engine**, which directs queries from data analytics methods towards the SIL and collects the results from the IoT platforms.
- The **Metadata Storage Component**, a database of metadata produced by the data analytics algorithms, which are necessary for their on-line operation.

In ACTIVAGE, the data collected by the IoT sensors and used for data analytics are stored in the storage facilities of each separate IoT platform. This facilitates the registration of new platforms, since it avoids switching to a different database and duplicating data. It also promotes data security and privacy, since the sensitive raw data remain in the deployment site's premises and under any site-specific privacy-related restrictions. However, the Data Lake does offer additional central storage, dedicated to metadata necessary for the operation of data analytics. These include produced features and analysis results, e.g. trained classification models, anomaly detection thresholds, etc., which may be necessary for their operation. Metadata are usually produced off-line, at regular intervals, using historical data, in order to be later used for real-time analytics.

The operation of the Data Lake and it connection to the SIL is described in Fig. 2. Data analytics methods (e.g. anomaly detection) need raw data stored in the distributed storages of the IoT platforms (e.g. the most recent sensor measurements), as well as specific metadata (e.g. pre-computed anomaly detection thresholds). The raw data are requested from the Data Integration Engine, while the metadata from the Metadata Storage Component. In order to collect the raw data, the Data Integration Engine submits a query to the SIL, written with the naming conventions of the unified ACTIVAGE ontology. The SIL translates the query to the platform-specific data models. The IoT platforms retrieve the requested data from their storage and return them to the SIL, which translates them to the ACTIVAGE ontology and sends them back to the Data Integration Engine. The latter combines the multiple sets of returned results and sends them to the data analytics component. At the same time, the Metadata Storage Component retrieves the requested metadata and sends them to the data analytics component as well. The data analytics method now has all the necessary information to produce the requested output (e.g. the detected anomalies).
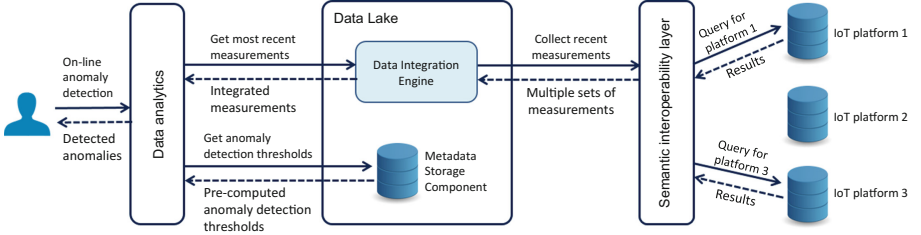
Fig. 2. Operation of the ACTIVAGE Data Lake.

3.3   Data Analytics and Information Visualization

The top layers in the ACTIVAGE architecture are the data analytics and information visualization layers, which provide meaningful representations of the raw data to the human operator. IoT applications are usually targeted at monitoring an environment, e.g. a person, a house, a city, etc, in order to facilitate decision making. In the context of e-health for older people, which is the primary target of the ACTIVAGE project, the purpose is to facilitate clinicians in monitoring an individual's health and taking proper actions, or to facilitate researchers in monitoring large sets of individuals and discover correlations. The focus of data analytics is thus on methods that extract representative features, find correlations, detect anomalies in usual behavior (e.g. to trigger alarms), and cluster objects (patients, devices, etc.) in groups of similar characteristics.

Existing data analytics methods are used in ACTIVAGE, covering the tasks outlined in Sect. 2: feature extraction, dimensionality reduction, anomaly detection and clustering. Table 1 summarizes the data analytics methods used in ACTIVAGE. This is not an exhaustive list, since other methods may be included as needed by IoT applications. Information visualization aims to produce descriptive graphical summaries of the raw data, allowing the operator to have a comprehensive overview of the data and explore them in order to detect interesting patterns. Table 1 summarizes the visualization methods used in ACTIVAGE. Commonly used visualization methods, such as bar charts and line plots are used, as well as more sophisticated graph-based visualizations for visualizing similarities and differences among objects.

4   Preliminary Evaluation

The proposed architecture is currently being evaluated using a smart home scenario and a smart mobility scenario. The purpose of the smart home scenario is to monitor the health status of older people as they perform activities of daily living, and assist the clinician in decision making through data analytics services. Environment and activity detection sensors are installed in the older person's home, constantly measuring temperature/humidity, CO levels, person motion and door/window opening. Two medical devices, a blood pressure monitor and a

Table 1. Data analytics and visualization methods used in ACTIVAGE.

Data analytics category	Methods
Feature extraction	PCA [9], MDS [10], Graph embedding [11]
Anomaly detection	SVM, Random Forests, LOF [6], BRPCA [7]
Clustering	Partition-based (k-means, k-medoids), hierarchical
Visualization category	Methods
Time-related visualizations	Line plot
Relative size comparison	Bar/pie/stacked bar/sunburst chart, treemap
Multi-variate comparison	Scatterplot, parallel coordinates, spider web, heatmap
Graph-based visualizations	k-partite graphs [35], multi-objective visualization [36]

blood glucose measurement device, are also used at specific times within the day. All devices are connected to the gateways via Bluetooth, ZigBee and ZWave protocols, while the universAAL [27] and IoTivity [20] platforms are used for their management. The scenario is currently being installed in testhomes, in order to be further deployed in several Greece municipalities, during the next period, with 500 scheduled participants in total. The purpose is to allow centralized management and analysis of the collected data by healthcare professionals.

In the mobility scenario, the purpose is to monitor and assist the older person while moving in a city, providing information and alerts when needed. The sensors involved include Bluetooth detectors installed at intersections for detecting bypassing devices, connected traffic signals, taxi data collectors, environmental pollutant detectors and pedestrian presence detectors. The FIWARE [25] IoT platform is being used for device and data management, with the aim to use more IoT platform types in the future, as part of a larger deployment. The scenario is currently being installed in test sites, with the purpose of being further deployed in Greece municipalities, with 500 scheduled participants. The purpose is to monitor the environment and the participants' movements, analyzing the collected data to provide notifications when certain patterns are detected.

5   Conclusion and Next Steps

This paper proposes an architecture for big data analytics in the IoT domain, in the context of large-scale federations of IoT platforms with heterogeneous data models. The semantic interoperability issue is addressed by introducing the Semantic Interoperability Layer (SIL), which maintains a common ontology describing relevant IoT concepts, as well as semantic mappings with the platform-specific ontologies. In this way, the upper layers can be agnostic of platform-specific naming conventions and semantics. The architecture also introduces the Data Lake layer, for directing external queries and results to and from the SIL, as well as for storing analysis metadata (extracted features, trained models, etc.) which are needed for real-time data analytics. The architecture is

being tested in laboratory environments, and is about to start being tested in real-world deployment sites. The architecture has been developed in the context of health assistance for older people, although it is generic enough to be applied in any application domain, such as smart cities, traffic monitoring, etc.

The next steps will be focused on implementation, integration and large-scale deployment. The proof-of-concept of the proposed architecture has been demonstrated in laboratory settings with a limited part of the whole architecture functioning. In the next period, the SIL ontology will be defined and implemented in detail, the Data Lake infrastructure will be completed to provide the basis for all data analytics methods, and the implementation of data analytics and visual analytics as Web services will be performed. In the meantime, integration issues will be resolved in order for the whole data analytics workflow to perform end-to-end. Finally, as mentioned in Sect. 4, the architecture is going to be tested in large-scale deployment sites in Greece municipalities, with a large number of participants, in order to use and evaluate it in real-world conditions. During evaluation, fine-tuning of ontology entities and data/visual analytics will be performed, in order to identify those concepts and methods that best fit in large-scale applications.

References

1. European project ACTIVAGE. http://www.activageproject.eu/
2. European project FrailSafe. https://frailsafe-project.eu/
3. Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I.A.T., Siddiqa, A., Yaqoob, I.: Big IOT data analytics: architecture, opportunities, and open research challenges. IEEE Access **5**, 5247–5261 (2017)
4. Soualhi, A., Medjaher, K., Zerhouni, N.: Bearing health monitoring based on Hilbert-Huang transform, support vector machine, and regression. IEEE Trans. Instrum. Measur. **64**(1), 52–62 (2015)
5. Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., Li, K.: A parallel random forest algorithm for big data in a spark cloud computing environment. IEEE Trans. Parallel Distrib. Syst. **28**(4), 919–933 (2017)
6. Papadopoulos, S., Drosou, A., Dimitriou, N., Abdelrahman, O.H., Gorbil, G., Tzovaras, D.: A BRPCA based approach for anomaly detection in mobile networks. In: Abdelrahman, O.H., Gelenbe, E., Gorbil, G., Lent, R. (eds.) Information Sciences and Systems 2015. LNEE, vol. 363, pp. 115–125. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-22635-4_10
7. Ding, X., He, L., Carin, L.: Bayesian robust principal component analysis. IEEE Trans. Image Process. **20**(12), 3419–3430 (2011)
8. Rokach, L., Maimon, O.: Clustering methods. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer, Boston (2005). https://doi.org/10.1007/0-387-25465-X_15
9. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometr. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)

10. Cox, T.F., Cox, M.A.: Multidimensional Scaling. CRC Press, Boca Raton (2000)
11. Yan, S., Xu, D., Zhang, B., Zhang, H.-J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. **29**(1), 40–51 (2007)
12. Kalpakis, K., Gada, D., Puttagunta, V.: Distance measures for effective clustering of ARIMA time-series. In: Proceedings IEEE International Conference on Data Mining, ICDM 2001, pp. 273–280. IEEE (2001)
13. Apache Spark. https://spark.apache.org/
14. Apache Hadoop. http://hadoop.apache.org/
15. Strohbach, M., Ziekow, H., Gazis, V., Akiva, N.: Towards a big data analytics framework for IoT and smart city applications. In: Xhafa, F., Barolli, L., Barolli, A., Papajorgji, P. (eds.) Modeling and Processing for Next-Generation Big-Data Technologies. MOST, vol. 4, pp. 257–282. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09177-8_11
16. Rathore, M.M., Ahmad, A., Paul, A., Rho, S.: Urban planning and building smart cities based on the internet of things using big data analytics. Comput. Netw. **101**, 63–80 (2016)
17. Bajaj, G., Agarwal, R., Singh, P., Georgantas, N., Issarny, V.: A study of existing Ontologies in the IOT-domain. arXiv preprint arXiv:1707.00112 (2017)
18. Compton, M., Barnaghi, P., Bermudez, L., GarcíA-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., et al.: The SSN ontology of the W3C semantic sensor network incubator group. Web Seman.: Sci. Serv. Agents World Wide Web **17**, 25–32 (2012)
19. Alaya, M.B., Medjiah, S., Monteil, T., Drira, K.: Toward semantic interoperability in oneM2M architecture. IEEE Commun. Mag. **53**(12), 35–41 (2015)
20. IoTivity. https://www.iotivity.org/
21. Open Connectivity Foundation. https://openconnectivity.org/
22. Soldatos, J., et al.: OpenIoT: open source internet-of-things in the cloud. In: Podnar Žarko, I., Pripužić, K., Serrano, M. (eds.) Interoperability and Open-Source Solutions for the Internet of Things. LNCS, vol. 9001, pp. 13–25. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16546-2_3
23. OPENIoT. http://www.openiot.eu/
24. Bermudez-Edo, M., Elsaleh, T., Barnaghi, P., Taylor, K.: IoT-Lite: a lightweight semantic model for the internet of things. In: 2016 International IEEE Conferences on UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld, pp. 90–97. IEEE (2016)
25. FIWARE. https://www.fiware.org/
26. sensiNact. https://projects.eclipse.org/proposals/eclipse-sensinact
27. universAAL. http://www.universaal.info/
28. Sofia2. http://sofia2.com/home_en.html
29. SENIORSome. http://www.seniorsome.com/
30. Veer, H., Wiles, A.: Achieving technical interoperability-the ETSI approach, European telecommunications standards institute (2008)
31. Daniele, L., den Hartog, F., Roes, J.: Created in close interaction with the industry: the smart appliances REFerence (SAREF) ontology. In: Cuel, R., Young, R. (eds.) FOMI 2015. LNBIP, vol. 225, pp. 100–112. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21545-7_9
32. Agarwal, R., Fernandez, D.G., Elsaleh, T., Gyrard, A., Lanza, J., Sanchez, L., Georgantas, N., Issarny, V.: Unified IOT ontology to enable interoperability and federation of testbeds. In: 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), pp. 70–75. IEEE (2016)

33. Noy, N.F., McGuinness, D.L., et al.: Ontology development 101: a guide to creating your first ontology (2001)
34. European project FIESTA-IoT: Federated Interoperable Semantic IoT Testbeds and Applications. http://fiesta-iot.eu/
35. Drosou, A., Kalamaras, I., Papadopoulos, S., Tzovaras, D.: An enhanced graph analytics platform (GAP) providing insight in big network data. J. Innov. Digital Ecosyst. **3**(2), 83–97 (2016)
36. Kalamaras, I., Drosou, A., Tzovaras, D.: Multi-objective optimization for multi-modal visualization. IEEE Trans. Multimedia **16**(5), 1460–1472 (2014)