The power of Open Data in the Internet of Thins

Federico Montori, Luca Bedogni
Department of Computer Science and Engineering (DISI)
University of Bologna, Italy
Email: {federico.montori2, luca.bedogni4}@unibo.it

***Abstract*—**

I. Introduction

Internet of Things (IoT) is one of the research and industrial fields that faced a rapid growth in the recent years. The approach to such ecosystem has been heterogeneous and sparse, leading to a wide variety of standards and solutions at each layer of the network and application stack. It has been a source of interest from many differnt points of view, in such a way that now we have dedicated infrastructures at physical, data link, network, transport and application layers.

Since the beginning of its diffusion, its potential has been explored in various fields of application and its major usefulness has been claimed to be in service composition [1].

The requirements when designing IoT-related automation systems are varying due to the heterogeneity of the platforms and the hardware components as well as the network interfaces. This resulted in a sparse set of technologies and terminologies used in several scenarios determining a lack of interoperability among systems. The common approach to the problem of unifying entities within an ecosystem is typically architectural and leads to a difficult reuse of the components among solutions [2]. To face these issues the European Commission supported initiatives like IoT-A [3], which aimed to release an architectural reference model and FI-WARE [4], which also helped architects in establishing an unified vision and nomenclature and now had become an implementation-driven open community. It also provided a sandbox in which partners could upload their open data, however it is not of broad use nowadays. ...

II. Related Work

The IoT is nowadays growing exponentially together with the number of solutions and architectures proposed to handle it.

[PREVISONI]

Regardless of the different perspectives, it is clear that the number of devices is growing, the data is becoming more and more heterogeneous, and one of the main challenges is how to handle such an amount of data and how to give a meaning to it. In the recent years there have been a huge number of attempts which, most of the times, are either self-contained since they require compliance to a specific framework, or commercial solutions.

[SOLUZIONI DI RICERCA][...]

[SOLUZIONI COMMERCIALI] Commercial solutions aim to constitute a living ecosystem in which entities are "plugged" and interoperable, participating for the benefit of the whole system and fully compliant with the other actors within the environment. Most of the times such frameworks, some of them depicted in [5], provide efficient software adapters for legacy systems. Such types of frameworks are often self contained and tend to create a cluster of devices which need to be framework-compatible in order to interoperate. [...]

III. Open data as a source

As stated in the introduction, open data are the most powerful source of information when such data are not producible by the utilizers themselves. In this section we outline some of the well-known sources that we considered in order to achieve a homogeneous data store.

ThingSpeak

ThingSpeak [6], originally launched in 2010 by ioBridge, is an open source data platform and API for the IoT that enables the user to collect, store and analyze data as well as interact with sensors and actuators easily. In more detail, it provides a personal cloud that users can deploy over their Local Area Network and easily display the data produced by sensors using ThingSpeak's straightforward API. Data analysis and visualization has been made possible due to the close relationship between ThingSpeak and Mathworks, Inc. since such functionalities are driven by the integrated MatLab support. Furthermore, such a platform provides a global cloud hosting millions of open data records, which is useful both to users who cannot deploy their own cloud and to consumers who need to infer information coming from the stored data. Data is stored with an absolute freedom of expression, meaning that any data record can have any name and it does not need to stick to any format constraint.

In the recent years, ThingSpeak had become very popular due to the rise of easily programmable IoT platforms such as Arduino, BeagleBone Black, ESP8266 and many others. Such devices are becoming cheaper and cheaper and, on the other hand, it is easier to get started with them. Nowadays, for instance, an ESP8266 is capable of manage a sensor, get connected through WiFi, be programmed through the simple,

C-like Arduino SDK and still cost less than 5$ while its battery, if the duty cycle is low enough, is estimated to have a duration of years. With a WiFi connection and an open platform such as ThingSpeak a first home sensor network is very easy to boostrap, since the device controller does not need to have the control on the cloud and, furthermore, the data produced by the sensor is easily displayable in a fancy way on the end consumer's personal device (a Smartphone or similar).

Sparkfun

SparkFun Electronics, Inc. [7], founded in 2003 in Colorado, is a microcontroller seller and manufacturer, known for releasing all the circuits and products as open-source hardware. It also provides tutorials, examples and classes.

For the purpose of the present paper, SparkFun also hosts its own open source cloud of open data [8], on which the customers can test and upload the data collected by the embedded sensors. Users can push for free their data on such cloud in streams of 50 MB maximum size and with a maximum frequency of 100 pushes every 15 minutes. Unlike ThingSpeak, the location where the data comes from is always specified at a coarse granularity since the name of the city is often obtainable, however the GPS coordinates are never given. On the other hand, data coming from SparkFun cannot be private.

Both SparkFun and ThingSpeak provide the data streams (or channels in ThingSpeak) using different markup notations: JSON, XML, CSV, MySQL, Atom and PostgreSQL. We extracted the whole repositories and parsed the JSON files in order to give a first structure to such data. Since the data structure does not force strong constraints data is, as explained in detail in section IV, often incomplete.

From each data stream we extracted the GPS position for a locational analysis, finding that such position is indicated, with different degree of precision, in XXX data channels out of XXX. In XXX% of the cases in which the position is specified, only a macro area is given (the city, or even the state). In such cases we took the central position of the indicated entity. The result of the analysis is outlined in figure 1.

From such results it is clear the importance of information fusion from different sources, since not only the sampling number of the sensing infrastructure is incremented, but also its coverage. Indeed, ThingSpeak appears to have much more utilization in the European region, whereas SparkFun seems to be more popular in North America. Furthermore, this consideration might be extended to different macro topic areas, meaning that some open data sources are specialized on a specific field of measuring. For instance, governmental sources providing open data such as EPA (United States Environmental Protection Agency) [9] are primarily focused on environmental data, whilst crowdsensing sources such as OpenSignal [10] regard measurements on cellular network signal strength and coverage.

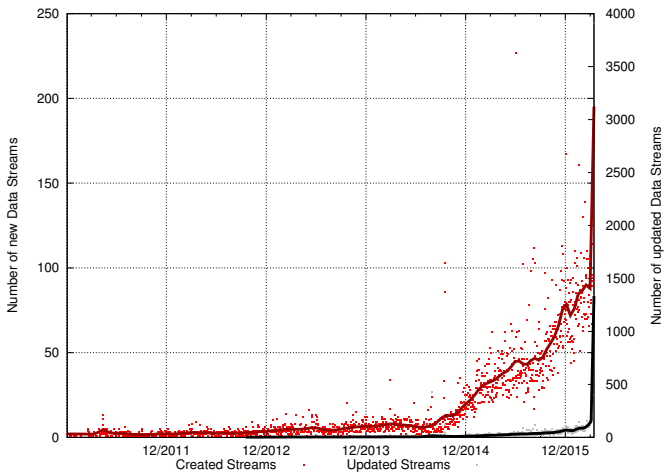

Fig. 2. Trend in creation of ThingSpeak channels.

Raw data streams also provide temporal information, especially regarding the creation date and the last update date. Such data can tell us much about the general trend in the usage of these platforms throughout a time window of few years. Each stream in ThingSpeak comes with a creation date, which we reported in the diagram in figure 2. From such analysis it results a substantial growth in created channels. Some of the steeper slopes are probably justifiable. A possible intuition behind them is the parallel innovation in simple hardware modules, for instance August 2014 corresponds to the launch of the first version of ESP8266 on the market and in October 2014 was possible to flash its firmware though an SDK [11]. Such period has not surprisingly been affected by a rapid growth according to the diagram. [MORE EXAMPLES] The diagram reports also the last update date for each stream. It is immediately clear that, since the oldest update date is in September 2012 while the oldest creation date is in 2010, all the data streams created before 2012 that did not perform an update after such date have been deleted due to a certain policy. The steepness of the curve in the last days reveals that a significant amount of the data streams are still in use and updated daily or even hourly.

Such analyses shed some light on how rapidly the world of open data is growing and people are gaining interest in using a platform that takes away the burden of creating a local ecosystem. Thus, our work creates the basis for a solid global ecosystem with a strong impact on cooperative services.

IV. Data Unification

In this section we point out the data streams' characteristics obtainable from our two open data clouds and how do we aim to unify them onto a single data cloud.

Hereby are briefly presented the parameters that can be extracted from data streams:

- **Stream ID**: it is the data stream's unique identifier. In ThingSpeak it is represented by an incremental number,

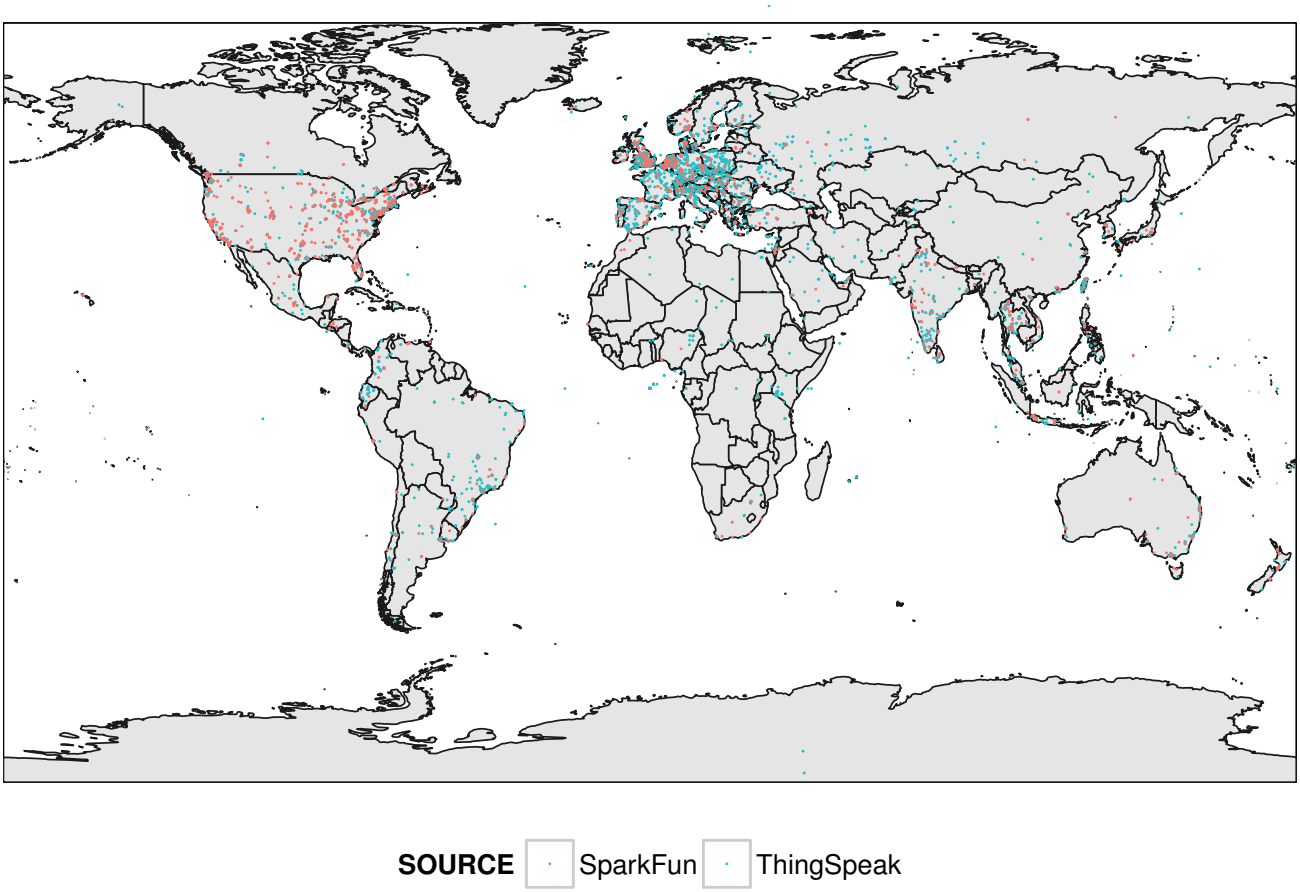

Fig. 1. Location of all ThingSpeak and SparkFun sensing sources.

which is assigned when the stream is created. On March 16, 2016 there are 28806 active streams with IDs spanning from 1 to 100172. In SparkFun the unique ID is given by a string of 20 random ASCII characters. On March 16, 2016 we counted 3575 different SparkFun IDs.

- **Stream name**: it present in both platforms and it is decided by the user with no constraint. This means that it can even carry no useful information for its identification and categorization.
- **Geolocalization**: it is present in both platforms. In ThingSpeak not all the streams come with geolocalization data, however it is always given in GPS coordinates. In SparkFun all the streams are geolocalized in the metadata, however there are no GPS coordinates, but only the name of the city, or sometimes just the state or even the country, unless the user indicates the GPS coordinates in the data itself.
- **Tags**: are included in both platforms and often help to infer useful information about the data.
- **Creation Timestamp**: it is included in all ThingSpeak streams as a metadata, on the other hand, the creation date of a SparkFun stream is not included. Each SparkFun stream is limited to 50 MB, thus the first update of any SparkFun stream less than 50 MB corresponds to its creation date, however it is retrievable in a limited number of cases.
- **Update Timestamp**: it is included in all ThingSpeak streams as a metadata. In SparkFun is simply deducible from the timestamp of the last update in the stream, since the timestamp is included in each data update.
- **Description**: it is a ThingSpeak metadata and its characterization is fully assigned to the user (who can also decide not to include it).
- **Elevation**: it is a ThingSpeak metadata and not always indicated.
- **Last Entry ID**: it is a ThingSpeak metadata, which points to the last update record in the data, ordered using an incremental ID for each update.

V. Conclusions

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] S. Krco, B. Pokric, and F. Carrez, "Designing iot architecture (s): A european perspective," in *Internet of Things (WF-IoT), 2014 IEEE World Forum on*. IEEE, 2014, pp. 79–84.
[3] "Iot-a european project," http://www.iot-a.eu/public/front-page, accessed: 2016-3-29.
[4] "Fi-ware open community," https://www.fiware.org, accessed: 2016-3-29.
[5] H. Derhamy, J. Eliasson, J. Delsing, and P. Priller, "A survey of commercial frameworks for the internet of things," in *Emerging Technologies & Factory Automation (ETFA), 2015 IEEE 20th Conference on*. IEEE, 2015, pp. 1–8.
[6] "Thingspeak, the open data platform for the internet of things," https://thingspeak.com/, accessed: 2016-3-29.
[7] "Sparkfun electronics," https://www.sparkfun.com/, accessed: 2016-3-29.
[8] "Sparkfun open data cloud," https://data.sparkfun.com/, accessed: 2016-3-29.
[9] "United states environmental protection agency," https://www3.epa.gov/, accessed: 2016-3-29.
[10] "Open signal," http://opensignal.com/, accessed: 2016-3-29.
[11] "Espressif sdk releases," http://bbs.espressif.com/viewforum.php?f=46, accessed: 2016-3-29.