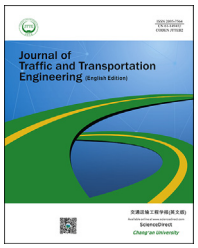
Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/jtte



Original Research Paper

Bike sharing systems data interoperability by a unified station status concept and big data solutions

Francisco Márquez-Saldaña [a,*], *Gonzalo A. Aranda-Corral* [b], *Joaquín Borrego-Díaz* [a]

[a] *Departamento de Ciencias de la Computación e Inteligencia Artificial, University of Seville, Sevilla 41012, Spain*
[b] *Departamento de Tecnologias de la Información, University of Huelva, Huelva 21007, Spain*

HIGHLIGHTS

- A unification methodology for storing bike sharing system data.
- A multi-system extractor and storage architecture for historical information.
- An efficient solution for big-data issues when storing bike sharing system data.
- A data serving toolbox from both machine-to-machine and human-computer.

ARTICLE INFO

Article history:
Received 5 October 2023
Received in revised form
23 May 2024
Accepted 20 June 2024
Available online 2 April 2025

Keywords:
Data acquisition
Big data in mobility
Bike sharing platforms
ETL

ABSTRACT

The impact of bike sharing systems (BSS) on urban mobility, and their study as part of the overall transport system in smart cities, has attracted significant academic interest in recent years. However, the lack of historical and standardized data in current service tools hinders the analysis and improvement of these platforms, i.e. by reusing technical data-based solutions. Big data nature (in volume, variety and velocity) of collecting BSS historical information must be also addressed, in order to take an integrated perspective.

This paper describes an integrated solution to this challenge by (1) proposing a unified station status concept for recording historical information, based on the identification, study and unification of common relevant fields found in almost all BSS data warehouses, and (2) implementing a big data-inspired ETL infrastructure together with a storage optimization, methodology which not only allows to access and collect previous defined concepts but also overcomes existing big data challenge when storing BSS information. The system also consumes other external relevant information, such as weather factors, which have been aggregated, enhancing stored knowledge, with KPIs and statistics. The developed solution illustrates how it can manage over seven years of data from twenty-seven BSS, serving not only machine-to-machine communication but also human-computer communication and enabling data-driven solutions.

* Corresponding author.
E-mail addresses: marquezsaldagna@gmail.com (F. Márquez-Saldaña), gonzalo.aranda@dti.uhu.es (G.A. Aranda-Corral), jborrego@us.es (J. Borrego-Díaz).
Peer review under responsibility of Chang'an University.
https://doi.org/10.1016/j.jtte.2024.06.003
2095-7564/© 2025 Chang'an University. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Modern cities generate and collect a wealth of information, giving rise to the field of urban data analytics, which has a notable influence on the modern reshaping of cities (Andrienko et al., 2021). Aggregating these diverse data streams enables improved processes and services, forming the basis for the concept of a "smart city" as well as the notion of its "digital twin" (Caldarelli et al., 2023). Data sources include dynamic data from sensors, as well as more static records such as national censuses, administrative and commercial databases, and even social and demographic surveys, which are less time-dependent (Borrego-Díaz et al., 2012).

Cities are complex systems with intertwined physical, social, and virtual networks that follow scalability and spatial principles (Batty, 2009). Nevertheless, the social behavior of citizens through their interactions with urban spaces and infrastructure most strongly determines interactions and dynamics that deserves in-depth study.

On the one hand when people feel a sense of belonging or purpose in a space, it impacts how they traverse and use the city as a whole. On the other hand, spaces that do not foster social connections or even feel unwelcoming tend to see little use or traffic, isolating them from the broader flow of people, resources, and activity in the city.

By drawing these together, useful systems can be built to improve city living. Whereas other branches of pervasive computing focus on specific technologies like wireless sensors, urban computing adopts a human-centered approach, considering how to enhance the urban experience holistically given the many data sources created and used within cities (Valle et al., 2010).

An aim of urban computing is to analyze cities in today's data-driven world through the geospatial cyber infrastructure (GCI) principles and tools. That enables capabilities like geospatial analysis and decision making (Yang et al., 2010). Analyzing a city's GCI is essential not only for developing a research agenda but also for understanding how a city relates to the physical, social, and geographical aspects of its environment.

On top of that, sustainability goals are driving efforts in urban areas, encouraging projects at national and international levels (Morton et al., 2017). Smart cities constitute global sociotechnological systems faced with sustainability challenges (Makarova et al., 2017). Information technologies, especially real-time tools and data, can empower individuals and groups to make continuous meaningful changes in response to constantly shifting circumstances (Millett and Estrin, 2012). As BSS are considered sustainable transportation, behaviour data analysis could serve to estimate its help in fields such as carbon footprint reduction, that comes from urban mobility. Hence, making it a necessary challenge for smart city projects (Makarova et al., 2017).

Moreover, BSS have supposed a game change in people's urban mobility behaviour during last years. Since the understanding of its role in city mobility has become a key, multiple approaches, involving a wide range of problems and challenges exists (Ricci, 2015). Its impact, together with the need of developing decision-making process and performing data analysis techniques for its study, has caught the interest of Academia.

The main goal of this work is to describe a solution to the aforementioned problems, through a big data-inspired platform, designed by the authors. It focuses on the rationale behind solutions for unifying, storing, and serving historical BSS data for a global analysis perspective, as well as the developed interfaces. The enterprise involves solving interoperability issues. A general station status concept is defined which accounts for historical records, a critical need for global analysis. Moreover, an efficient extraction methodology must be implemented (the requisite extraction-transformation-loading or ETL phase for data science projects). Notice that this work substantially extends previous communication (Marquez-Saldaña et al., 2022).

To enhance knowledge extraction process, external related data, such as weather factors, could be highly interesting. Another goal would be to publish recorded information for supporting tasks from researching to monitoring system health, passing through just checking station information by affiliate users. In the end, providing new information windows for the so-called smart city control rooms (Marazzini et al., 2018).

To accomplish aforementioned goals, available technologies designed for ETL and data service solutions have been analyzed. Finally, those which fit problem requirements the best has been merged in the final solution which could be considered a whole cycle in BSS industry taking into account sustainability objectives of smart cities projects. Moreover, all developed solutions for this work are under open data common attribution (ODC-BY1) and creative commons (CC-BY 2.02) license.

This paper is organized as follows. First part (section 2), focuses on a literature review of existing studies about BSS from a data-driven perspective. Then, in section 3 there is presented a unified station status and collector methodology as a path for BSS data interoperability. Next, in section 4, there are presented three developed consumption tools for covering both sides of data acquisition tasks (computer-computer and human-machine communication). Potential uses-cases and applications are described in sections 5 and 6 respectively. Finally, conclusion and general considerations on the system are discussed together with future work in section 7.

2. Related work

The interest of researchers in analyzing BSS has been followed by two well defined paths. On the one side, integrating, storing and making available BSS data could be considered one of the first challenges to be faced. On the other side, extracting knowledge from recorded information could drive to a better understanding of BSS and its impact in cities' behaviour.

2.1. *ETL and big data issues in bike sharing systems*

The global nature of bike sharing systems enables comparing solutions across diverse cities and contexts. To analyze these

systems, institutional leaders, urban planners, and smart city experts must first compare system behaviors and data. Studies of individual cities, like ongoing work of Froehlich et al. (2009), cannot derive best practices from other cities' performance data. Addressing the variety and volume of BSS data is essential (Andrienko et al., 2021), both for docked and dockless systems (Costa and Silvestri, 2021). Also, data generation velocity (both generation and consumption by data driven systems) must be considered. In summary, storing and serving BSS data requires to face a problem with common characteristics to big data projects.

While big data tools and frameworks exist to manage BSS information (Jia et al., 2017), the lack of conventions remains a barrier to gaining knowledge and insights into universal behavioral patterns across systems. This issue has arisen due to the significantly increase of existing BSS in the last decade without taking into consideration standardization practices. Thus, it impedes extracting knowledge and identifying universal behavior patterns across systems.

Several approaches looking for the standardization are available. On the one hand, the proposal general bikeshare feed specification (GBFS, https://github.com/NABSA/gbfs) represents a good methodology for storing BSS data, recording historical information is not considered. On the other hand, the so-called mobility data specification (MDS, https://www.openmobilityfoundation.org/about-mds/) based on the previous one, takes into consideration the importance of historical registers and even merging BSS records with other urban mobility data. However, it is not open data oriented as it is only available for commercial companies, mobility regulators and public agencies.

All of that compels researchers to face problems of analyzing BSS through a data perspective, increasing the difficulty when comparing more than two systems. Despite efforts focused on data interoperability primarily by the open data community (Charalabidis et al., 2018), the commercial nature of municipal contracts does not promote these practices. These factors, combined with a lack of historical data necessitate a unified, historical BSS data repository.

2.2. *Extracting knowledge from BSS data*

The huge amount of BSS available information, together with the wide range of analytical tools developed in recent years, have enabled BSS issues being faced through a data-driven perspective. One of the motivations behind this work is that the analysis and BD infrastructure developed herein can be beneficial for proactive approaches to the design of bike-sharing networks, aiding in data processing to understand dynamics, optimize services, and forecast growth.

First, understanding the impact of BSS in cities and how it has changed urban mobility has been tackled. In the study of Ricci (2015), there is presented several benefits of BSS related to users' health improvement, and commute time or cost reduction. However, there is mentioned that other relevant theoretical advantages such as traffic congestion, and hence, pollution level, reduction requires an upgrade in existing ETL solutions applied to BSS. Moreover, to achieve previous benefits it is also required to analyze BSS usage variations among both, geographical and sociological circumstances (Pearson et al., 2022). In relation to that, security is one of the main aspects which influences in users number. Hence, city infrastructure relation with cycling accidents has been studied by Daraei et al. (2021).

Another study focused on BSS impact (Natera Orozco et al., 2020), where the authors have identified multiplex profiles based on transportation data from 15 cities. The study finds that most cities have fragmented bicycle layers (pedestrian, bicycle, rail, and street) and proposes strategies to unify these disconnected components.

Other frequent analyzed field is BSS station structure improvement, normally considered as a network optimization problem. For instance, in the paper of Szell et al. (2021), models are developed to efficiently grow synthetic bicycle networks in 62 cities starting from street networks and points of interest, utilizing three growth strategies: betweenness, closeness, and random. In addition, the book of Vogel (2016) presents a complete review of data-driven methodologies for BSS station grid design. Moreover, in the study of Reggiani et al. (2022), the novel concept of "bikeability curves" is defined as a better metric of system's topology goodness. On top of that, the influence of topology and subscription price in BSS usage has been pointed out (Jurdak, 2013). Apart from station placement, there has been also developed an optimization framework to guide bike lane planning based on BSS trajectory data (Ferenchak, 2023; Liu et al., 2022b).

Exogenous factors have been also considered relevant for a complete understanding of BSS. On the one hand, city size and population density are directly related with system usage (Jiménez and Nogal, 2021). On the other hand, a statistical study across several climate zones (Bean et al., 2021) shows that, behind weekday, weather factors such as precipitation and high temperatures has been pointed to influence the most in BSS daily trip count. Furthermore, climate conditions impact has been also confirmed through a random forest model (Ashqar et al., 2019).

BSS data quality improvement is also a wide developed path. An outlier detection methodology based on functional analysis (Liu et al., 2022a). Moreover, clustering techniques for identifying anomalies and improve bike demand forecast accuracy has been implemented in Rennie et al. (2022). Furthermore, there has been pointed that deep learning classification techniques fed with trajectory data could be also applied to identify anomalies in cycling behaviour (Yaqoob et al., 2023).

On top of that, analytical solutions are also developed for solving BSS well known issues such as the static unbalancing problem (Dell'Amico et al., 2014). Therefore, several approaches based on Tabu search (Chemla et al., 2013), the branch-and-cut algorithm used for the "one commodity pickup and delivery traveling salesman problem" (Erdogan et al., 2014) or a continuous time Markov chain implementation (Federico et al., 2018). From a different perspective, unbalancing problem could be analyzed by using multiagent modelling and Q-learning (Shimizu et al., 2013).

Bike movement forecast solutions could be found at several levels of complexity, depending on requested precision and spatiotemporal scope. Thus, simple models such as Bayesian networks present enough performance for short

periods without stationary behaviour (Froehlich et al., 2009). On the contrary, predicting for long periods and handling with seasonal patterns at station level could be solved by using more complex methods such as recurrent neural networks combined with long sort term memory (RNN-LSTM) (Lim et al., 2022) or temporal convolutional network plus gated recurrent unit (TCN-GRU) (Li and Xu, 2024; Zhou et al., 2022).

Finally, proper ETL processing of the data (and its initial analysis) can be utilized to construct simulation models or to estimate essential parameters not only applied for BSS, but also for transportation generally. For instance, similar to the type presented in the paper of Barbet et al. (2021) (where an agent-based model designed to examine the modal shift in public transport following disruptions is presented), but with a focus on integrating bike-sharing services. The work of Li et al. (2017) proposes nonlinear solution to calculate relative arrival rates, based on the product-form solution for the stationary probabilities of joint queue lengths at virtual nodes, is proposed. In the study of Li et al. (2024), authors introduce a novel framework named DualST for predicting urban flow, which uniquely separates temporal semantics into closeness and periodicity maps to better understand temporal patterns and dynamic trends.

3. Unified data collector

The absence of semantic commitments (e.g., an ontology for properly specifying metadata), together with the lack of conventions in BSS data repositories, obstructs data science-based services. Variable notation and similar conceptual discrepancies between multiple data storage systems (semantic heterogeneity) need to be addressed. Let us focus on some of the critical discrepancies for this project.

First of all, the meaning of available stands is not interoperable, which could include not working stands or just be an aggregation of available bikes or docks depending on the source.

Concerning update time frequency, relevant diversity is observed even between the same data access point, which presents analysis and comparison tasks limitations. This problem worsens when information must be aggregated with external data, such as weather variables, as it comes from a different origin with no fixed time range neither.

Additionally, most original BSS data sources only grant access to the current system, making historical records unavailable. The limited access increases the velocity requirements of the extraction solution as the information acquisition process must run continuously to avoid data loss (e.g., every 5 min).

All these issues will be discussed below and have been faced by the implementation of the extraction system presented in Fig. 1.

3.1. *Unified station status concept*

By designing a unified specification for station information focused only on those concepts necessary to represent station evolution over time, the aforementioned syntactic and semantic heterogeneity among data sources has been addressed. To achieve this goal, data delivery solutions from multiple system owners were analyzed, such as JCDecaux (https://developer.jcdecaux.com/#/home) or CitiBike (https://citibikenyc.com/system-data).

From the above analysis, the number of available bikes, empty docks and total stands have been chosen as the main concepts for describing a station at a given time. Note that the last is required so as not to lose information on broken stands, which is quite useful in quality of service (QoS) analyses.

Therefore, the time-dependent unified station status (USS) at a certain time will be specified by the following quadruple.

- **Timestamp.** To keep historical order.
- **Available bikes.** Bikes that can be taken without taking into consideration its kind (mechanical, electrical, etc.).

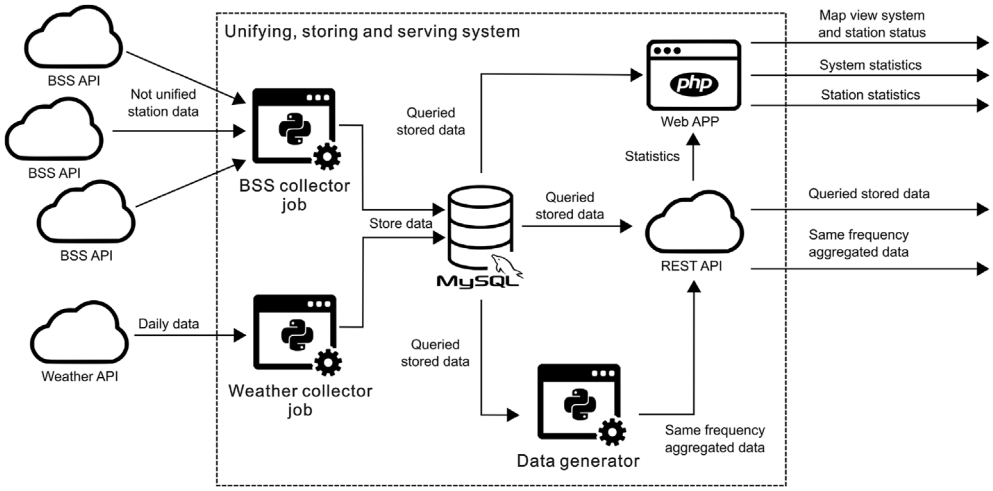


Fig. 1 – Functional diagram of the entire system. Note: data are retrieved by collectors and integrated in a centralized database. Both external and internal access is handled by developed consumption tools.

- **Available docks.** As the number of empty stands for returning a bike.
- **Total stands.** The overall station capacity, including not working stands.

Historical evolution of a system could be defined as a set of station statuses, which includes at least the USS criteria through time. To enrich this information, other system domain-specific features can be stored whenever no notation and meaning discrepancy with previously registered metrics is granted. Those extra metrics cannot subtract information to the USS. For example, electrical bike availability could be stored only if they are also taken into account in available bikes.

3.2. *Data extraction process*

On the one hand, regarding data nature, the set of historical USS stored has to be defined as dynamic data. On the other hand, as the final solution is intended to include and aggregate records from multiple sources, all system and station-relevant metadata are considered static information for the developed extraction tool.

Concerning system description it is required to store at least source identifiers (numerical or descriptive), country and time zone for handling multi-time location comparisons. In relation to stations, apart from identifiers it is quite interesting its geographical coordinates for identifying movement patterns. With respect to the key dimensions of the project, the following decisions have been made.

Variety: the data extraction process was designed using an active Python solution due to the variety of tools it offers for handling databases, data structures, and APIs. The developed extractor can not only read and process original information but also adapt and store it in the defined unified structure. As each BSS could be served through different APIs, one extractor must be implemented for each endpoint. To address this variety issue, the complete extraction solution is considered as the set of individual ones.

Velocity: to determine execution frequency, around 2500 station update times were analyzed to strike a balance between avoiding information loss and not overloading server storage and computing capabilities. The full extraction process launches every 5 min. This frequency is considerably lower than the regular update time of all stations (Fig. 2).

Volume: the extractor will insert a station status only if there is any change compared to the previously saved record. This represents an important storage efficiency improvement, reducing the total storage size by more than 10 times (from approximately 2.8 billion to 220 million). This strategy also avoids both information loss and duplication.

Finally, though static data is not expected to change, the implemented extraction logic can detect and update database metadata when needed. In addition, since some variations may affect critical fields used by extractors, such as commercial names or source identifiers, the solution also warns the administrator in those cases.

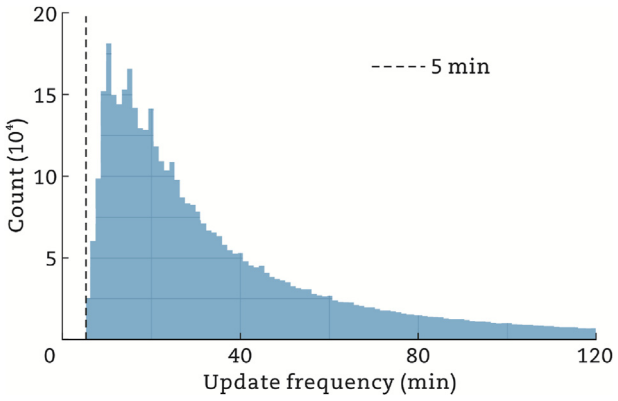


Fig. 2 – Update time frequency distribution.

3.3. *External data*

In relation to external factors, one that influences BSS behaviour the most is weather (Ashqar et al., 2019; Sanmiguel-Rodríguez and Giráldez, 2019). Thus, it has been consistently aggregated in the implemented system.

After analyzing several available alternatives, the government NOAA weather API (https://www.weather.gov/documentation/services-web-api) has been chosen as the weather data source. This adheres to the open-data philosophy, providing worldwide information that simplifies the extraction process for both current and future stations. As higher-frequency data are not under an open-data license, daily values for air temperature, precipitation, snow, and wind are saved.

3.4. *Data storage architecture*

All extracted information is stored in a centralized resource to simplify future data transformation processes. To find the most efficient database engine for this purpose, MySQL and MongoDB were analyzed, covering the most used solutions in the current data architecture paradigm (SQL and NoSQL, respectively). Both alternatives follow open-source criteria, as does the project. While emergent data storage methodologies such as data lakes could be considered, current server hardware limitations make them infeasible; thus, they were not included in this study.

To decide between previously mentioned technologies, the methodology has involved performing query time and storage space charge test on them. These metrics could be thought as good estimators for real-time data consumption when handling big-data solutions. In consequence, it has been loaded one year of data of an over the mean size system (260 stations) resulting in more than four million records.

Concerning query time performance (Fig. 3), although both engines consumption query time increase lineally, MySQL slope is considerably lower than presented by MongoDB, resulting near four time faster in the biggest query. With

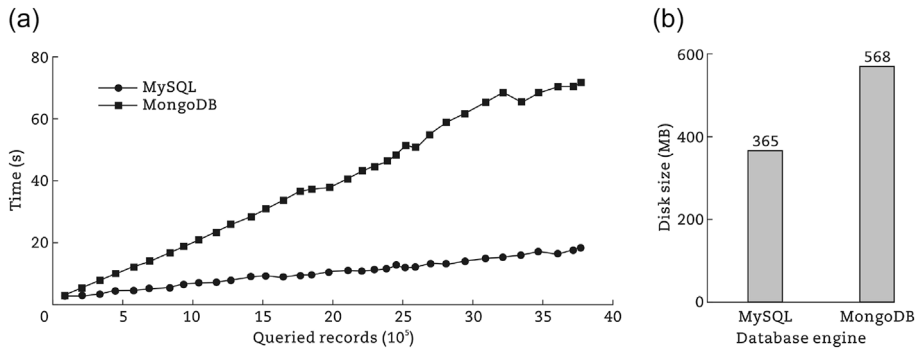


Fig. 3 – Database engines time and storage comparison. (a) Query time efficiency. (b) Storage efficiency.

regards to disk space, MySQL data comprehension overcomes again MongoDB famous BSON format (365 against 568 MB) when storing BSS information.

Consequently, as it presents the best perform in all developed test, MySQL has been chosen as the database engine for this work.

3.5. *Database schema*

Two tables were defined for storing both system and station metadata (Fig. 4). To avoid server hardware bottlenecks when querying large tables, a separate one containing all USS is created for each system regarding status history. Therefore, apart from pointing to the separation between static and dynamic data, this design reduces storage space for large systems. In addition, isolating queries at the system level maintains computing performance from infrequently updated systems to highly frequently updated systems increase.

To ensure BSS comparison being possible, in each status table must be included at least the previously defined USS in this work. For enriching stored content with other relevant domain-specific data, a new adaptation becomes necessary to maximize unification.

Concerning external data, it is stored in a single table for all systems, because selected daily frequency does not impact in storage efficiency. In addition, as information source normally includes one weather station per system, there is no necessity to store values per each BSS station, aggregating in those cases where more than one metric might be available for the same day.

Despite the defined storage architecture prevents data duplicity and reduces queries to retrieve data, recorded timestamps do not match even between stations of the same system. This issue implies significant difficulty for comparison tasks, which is resolved not only through information schema but also by the data consumption tools developed in this work.

4. Data consumption tools

For making unified data available, several tools for data processing and both human and machine communication services have been developed. They not only solve several previously mentioned problems concerning data reduction; they also enrich the information of delivered value.

4.1. *Data generator*

Aforementioned unmatched recorded timestamp problem of system or station comparison process is faced through a Python tool. The solution presented in this section is able to generate same-frequency status data from stored records within a date period requested by the user. Also, returned step frequency can also be defined to cover future analyses requirements.

Recalling that to reduce volume, data are not stored by extractors until a change occurs, even if the status remains the same for several hours as the last recorded one. This tool is based on that principle. It basically consists of three steps. First, interpolated data are obtained by querying recorded statuses between the requested period. Second, same-step window interval timestamps are constructed. Finally, the previous actual status is matched to each generated one. To illustrate the process, Fig. 5 shows an example of a 15-min interval user data requests from 10:30:00 to 12:00:00.

Fig. 5 illustrates how data generator tool matches each status to a 15 min interval, repeating the same value to fill the gaps when necessary. Notice that same color is used for representing original stored record (left side) translated to generated information (right side). Status data is represented as the defined triple: available bikes, available docks, total stands.

The tool is designed not only to aggregate BSS data from different sources but also to merge with other relevant external information for the system, such as recorded weather metrics. Accordingly, this generator can be enhanced by adding new metrics supplying demand of other solutions presented in this work.

It is worth to mention that as the tool is designed as an inner part of the system, that is, it cannot be directly accessed by the final user. Instead, these options will be included in the final developed RestFul API.

4.2. *API rest*

For enabling automated information and knowledge extraction from the presented system in this work, a Python RestFul

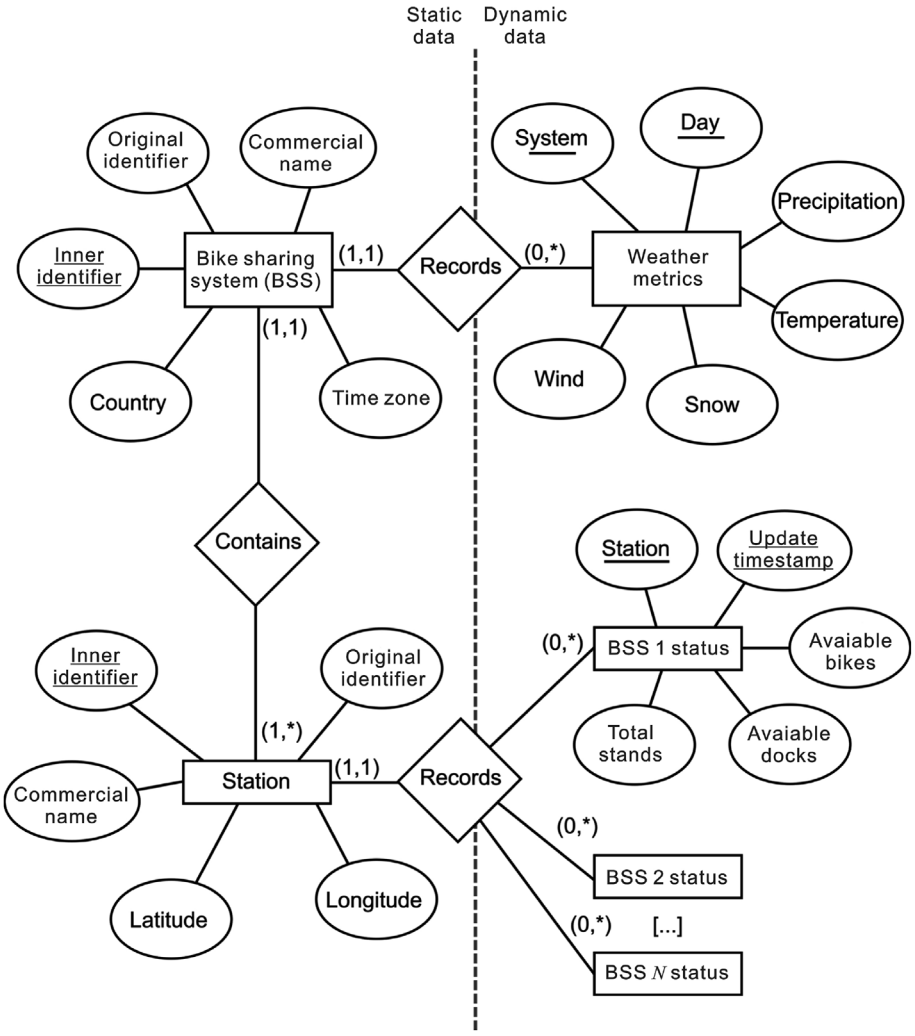


Fig. 4 – Database entity relation diagram.

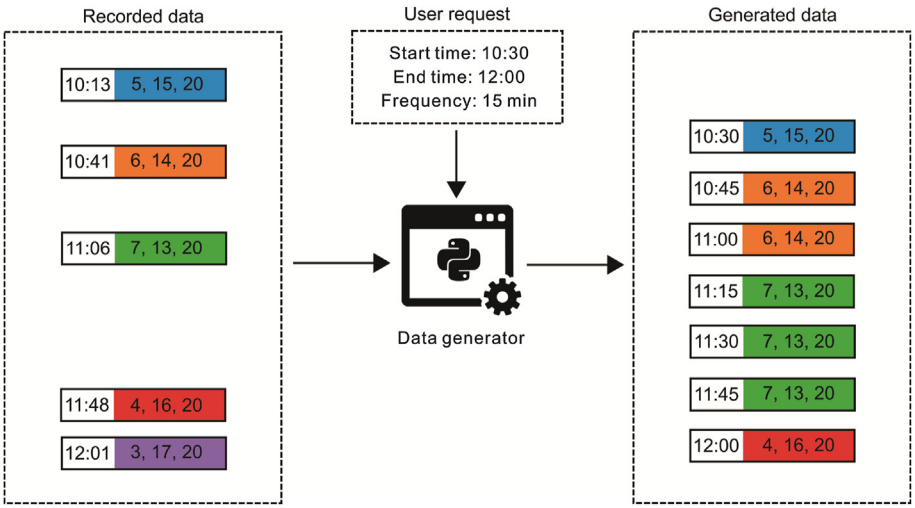


Fig. 5 – Data generator workflow for an example user request.

Call: /data/systems/<system id >/station/<station id>/status?start time=YYYYmmddTHHMMSS &end time=YYYYmmddTHHMMSS&frequency=1h
Response:
[
{"source_id": "1","commercial_name": "001_GLORIETA OLIMPICA",
"latitude": 37.4129235511391,"longitude": -5.98890593824315
},
{"source_id": "2","commercial_name": "002_GRAN PLAZA",
"latitude": 37.381578045327934,"longitude": -5.96522396639778
},
{"source_id": "3","commercial_name": "003_PUERTA DE LA BARQUETA",
"latitude": 37.40564154237,"longitude": -5.99848824083126
},
{"source_id": "298","commercial_name": "298_LAB MADRID",
"latitude": 0.0,"longitude": 0.0
}
]

Fig. 6 – Simple API call request example and API response used for retrieving recorded station metadata for Seville.

API has been developed. It has been developed using Flask Library as its lightness reduces call time. In relation to naming patterns, commonly accepted convention (https://restfulapi.net/resource-naming/) has been followed. Consequently, status data is accessed hierarchically from systems to stations defined sources. Data filters are passed as HTTP parameters inside the request.

Concerning usage limits, whereas different system meta-data can be extracted at once, status information must be accessed station by station due to server limitations. That makes necessary to iterate over them to extract full system status. Finally, returned data will be in JSON format as commonly used nowadays, for supporting unification.

As a result, at the time of publication of this work, API provides actions such as getting meta-data for both systems and stations. It must be mentioned that this API is in experimental status at the publication of this study assuming that it could be future changes in the future or it may present unexpected behaviour. In Fig. 6, there is an example of a simple api call for requesting station metadata.

Concerning USS, it is possible to retrieve single status (at any recorded time) of a station or extracting both real records and same-frequency statuses between two dates. Notice that in the last one is where developed data generator takes place. An example of how to get hourly sampled status evolution of a station is shown in Fig. 7.

Regarding external information, it must be accessed separately by the system, without date restriction as only daily summaries are returned. Then, all recorded metrics for the requested date period are delivered in total.

The API also serves as an engine for the Web Application developed in this work. This means that the service provides all information needed by status maps and statistics dashboard implemented. It must be mentioned that these records cannot be directly accessed out of the system as this information is thought to be consulted through the Web Application.

Therefore, the implemented API is the heart of data flow for both external and internal processes.

4.3. *Web application*

Apart from enabling pragmatic access to stored information through the implemented API, a web application (http://opendatalab.uhu.es/bikes/) has been designed as a user-friendly access point. This solution includes relevant key performance indicators (KPIs), enriching with additional knowledge and reducing the effort required for future analyses.

For covering most needs, it has been implemented two main views. On the one hand, current system status with geolocation information. On the other hand, a deeper system and station behaviour knowledge for managing tasks or future research. This also draws a clear distinction between end-user types.

In relation to check current or historical status, it has been implemented a system map view (Fig. 8). Then, each station's

Call: /data/systems/<system id >/station/<station id>/status?start time=YYYYmmddTHHMMSS &end time=YYYYmmddTHHMMSS&frequency=1h
Response:
[
{"update_time": "2023-01-01T00:00:00","total_stands": 20.0,"available_docks": 16.0,"available_bikes": 1.0},
{"update_time": "2023-01-01T01:00:00","total_stands": 20.0,"available_docks": 15.0,"available_bikes": 2.0},
{"update_time": "2023-01-01T02:00:00","total_stands": 20.0,"available_docks": 15.0,"available_bikes": 2.0},
{"update_time": "2023-01-01T03:00:00","total_stands": 20.0,"available_docks": 14.0,"available_bikes": 3.0},
{"update_time": "2023-01-01T04:00:00","total_stands": 20.0,"available_docks": 14.0,"available_bikes": 3.0}
]

Fig. 7 – Complex API call request example and API response used for exploiting data generator capabilities. Note: more specifically, the call ask hourly statuses from 00:00 to 04:00 on January the first 2023 for a station located in Seville.

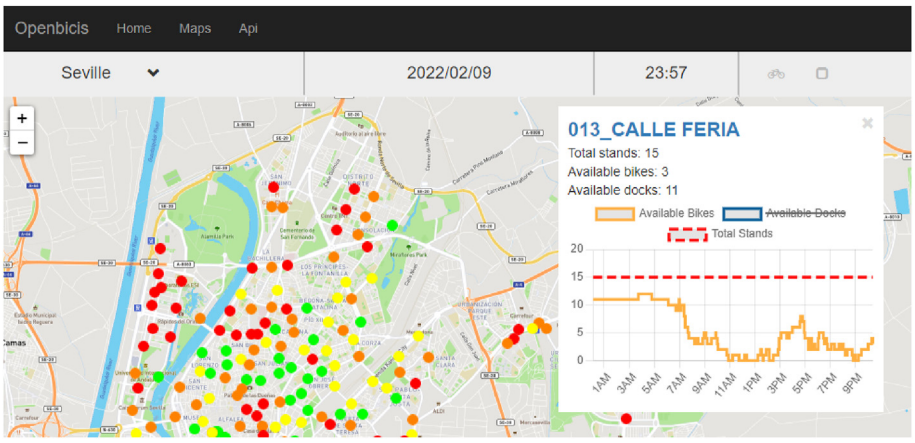


Fig. 8 – Map view shows the system status at a given time, displaying the status changes for each station over the course of that day.

bike or dock availability (selected by user) is presented as point colored from red to green, referring to less or more occupancy respectively. By clicking on a station, chosen daily metric trend is displayed, enriching consumer experience.

For implementing the map view it has been used existing open-source solutions, such as Open Street Map (https://www.openstreetmap.org/) and Leaflet.js (https://leafletjs.com/). They bring more useful functionalities apart from just displaying a map and follows open-source philosophy.

Regarding a more advanced usage, dashboards with relevant monthly and yearly statistics has been designed for both stations and systems on the whole. The aim of this is to bring more expert users a fast tool which supports analyses without programming calls to the developed API.

It has been selected those metrics commonly used or considered interesting for maintenance and researching. Such indicators are presented as charts of several types by using the well-known Javascript open-source library Chart.js (https://www.chartjs.org/).

Before explaining resulting dashboards, it must be mentioned that all metrics are based on the total status changes. As, in most of cases, it is not available bikes path tracking between stations, the previous criteria has been thought as the best indicator to analyze system and station usage.

For a general evolution overview, it has been designed a view with monthly information about system and station activity by weekday and hour (Fig. 9). Similarly, it is shown not working stands tendency, measured as total stands minus available bikes plus available docks. In addition, in station view it is also returned an extended monthly status evolution trend.

Concerning more specific statistics, in Station Dashboard it has been developed a heatmap which describes what hours the station is used the most for each weekday (Fig. 10). It has been constructed by overlapping all same weekday (for example, all Mondays) hourly usage and using transparency to accurately split real behaviour, as darken zones, from outlier records as lighten ones.

In relation to system specific metrics, it is considered to be useful to identify what zones of the system are used the most. Thus, it is included a station usage heatmap (Fig. 11) in which it is pointed each station in a map view and colored from cold to warm tones whether it has been used less or more

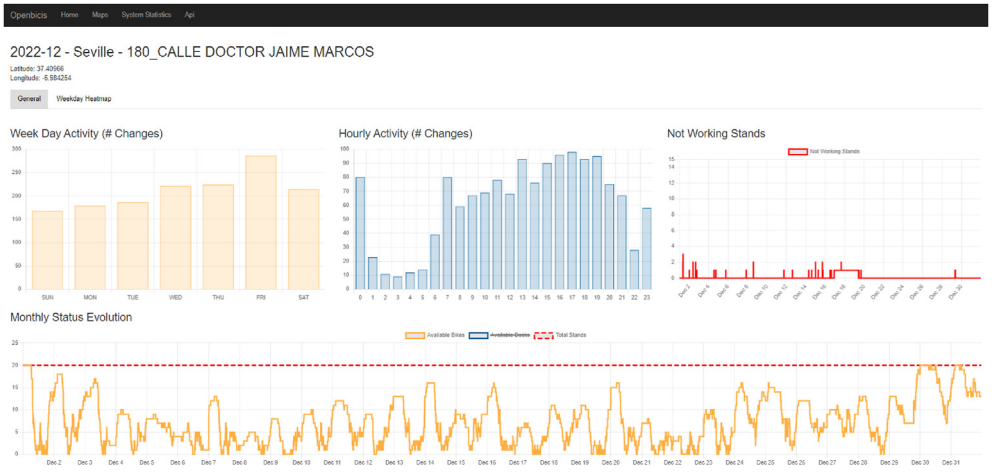


Fig. 9 – Chart of general station's KPIs and for one moth view.

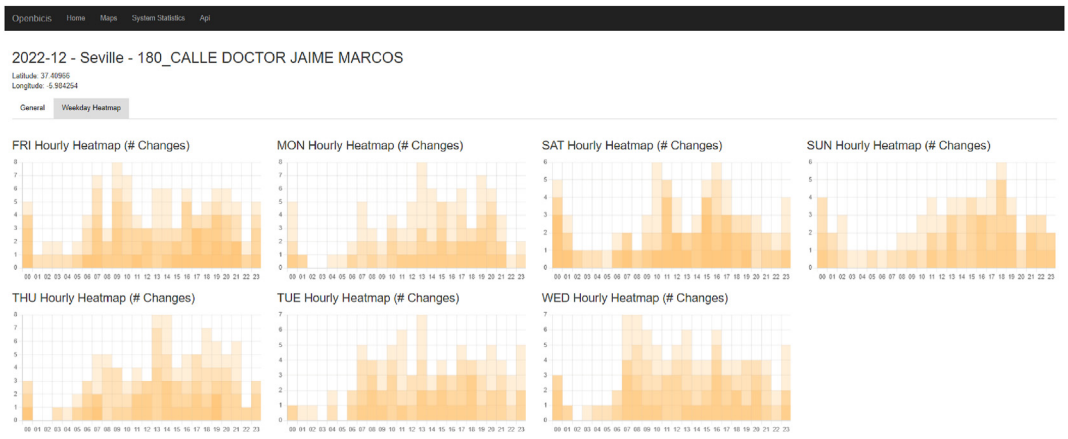


Fig. 10 – Weekday heatmap for one moth station view.

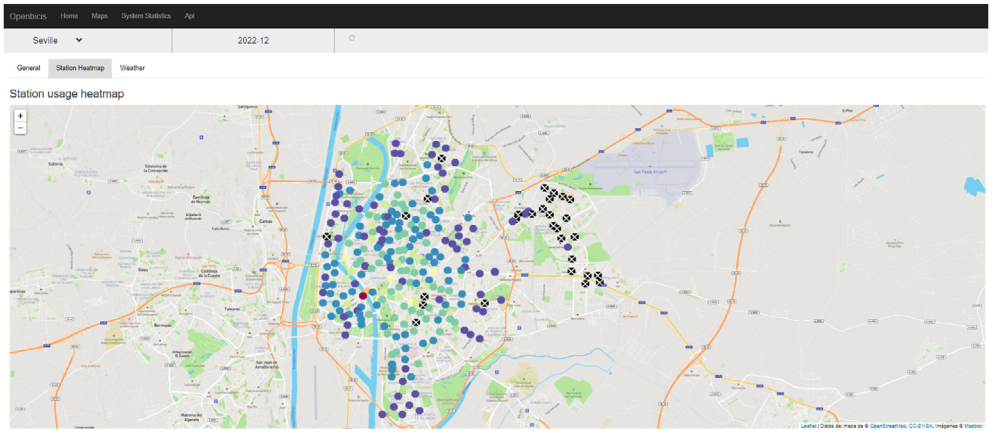


Fig. 11 – Station usage heatmap for system one moth view.

respectively. An extra dashboard, which shows a relation between system usage and several weather conditions recorded, is designed (Fig. 12). Notice that, due to strong seasonal influence on weather factors, the dashboard reflects yearly aggregations instead of monthly metrics to avoid patterns misunderstanding in features such as usage by temperature.

5. System user profiles and uses cases

Recording live and historical data from bike-sharing systems is useful across a wide range of areas and may interest diverse types of users (e.g., for monitoring or prediction). For that reason, this work aims to cover as much functionality as possible, not only for scientific purposes but also for improving bike-sharing systems and urban mobility experiences overall. As a result, four user profiles were specified for the tools presented in this work, as shown in Fig. 13.

5.1. *System-affiliated users*

Final customers must be considered as they are supposed to exploit the bike service. Relevant information available in presented application could be station bike or dock availability near both their start and end journey location. As a step

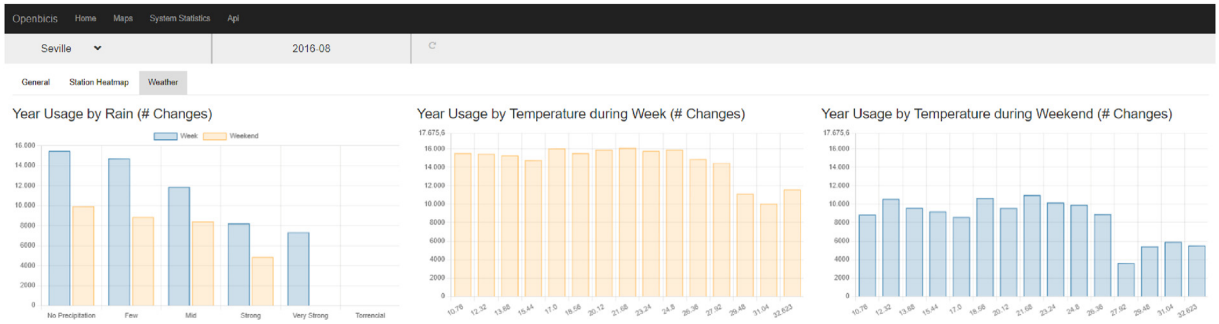


Fig. 12 – Yearly system usage by weather conditions.

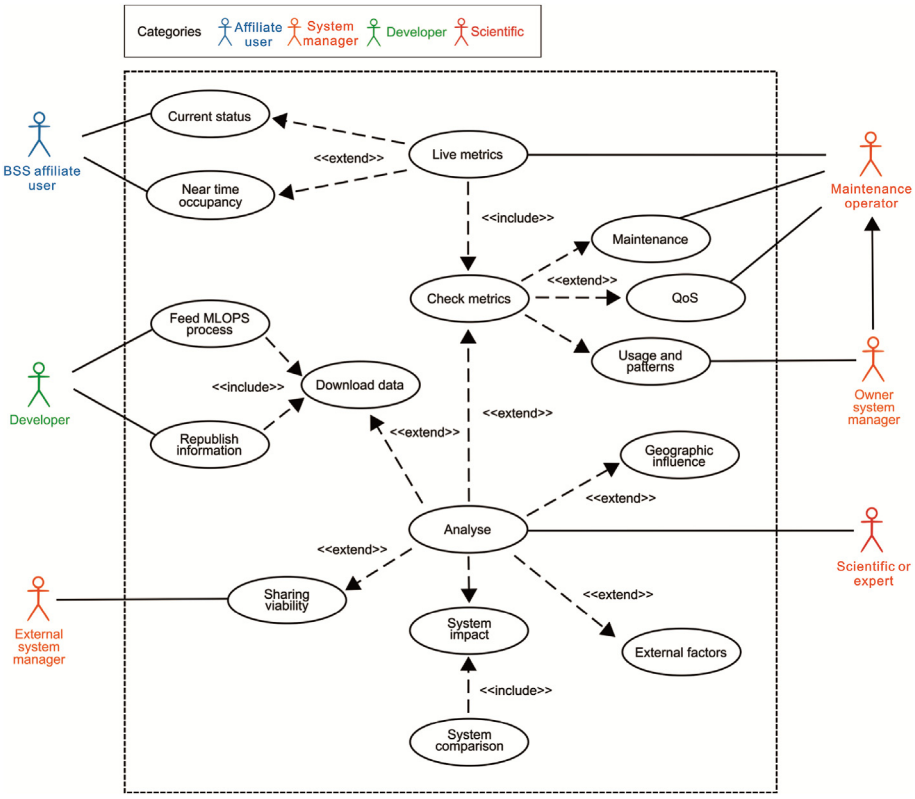


Fig. 13 – Full system uses cases.

further, they may find interesting to have an estimation of a station occupancy at a certain time (for example, each Monday at 09:00). Related to this, it could be used statistics pattern metrics previously defined in the application.

5.2. *System management users*

For users responsible for bike-sharing systems, the presented work could prove helpful. First, maintenance operators may find the developed dashboards valuable, with quality of service or unbalanced real-time station KPIs available in the application. System owners or managers could gain deeper insight into bike-sharing usage and patterns related today, time, weekday, season, geographic area, etc., enabling a better understanding of the system overall and its needs.

It should not be forgotten that other managers external to the system could use the recorded data and dashboards to assess the viability of "sharing" in other emerging areas such as scooter- or motorbike-sharing systems.

5.3. *Developers*

Other relevant actors could be automated (intelligent or software) agents used by developers. By downloading raw or processed data through the developed RestFul API previously mentioned, they may keep up to date with BSS information or even republishing and/or processing in third-party applications enhancing BSS knowledge. The system could be helpful for more advanced tasks, such as feeding with new data existing models for forecasting services.

5.4. *Scientific or expert*

Finally, more scientific or expert profiles (e.g., urbanists or smart city developers) must be considered given the importance of bike-sharing systems in their work. Seven years of historical data could help both traffic experts and urban mobility researchers analyze the impact of bike-sharing not only on other mobility modes but also within the context of smart cities. The wide range of locations from which BSS data are collected may enable comparing and studying the geographic influence of dock networks.

In addition, exogenous factors aggregated could contribute for a better understanding of BSS behaviour. Perhaps, by allowing to explain some patterns bringing more robustness to future studies.

6. A glimpse of potential applications

The primary aim of the platform is to provide unified and useable information for supporting data driven tools or studies. For example, simulations based on multi-agent systems (MAS) over pedestrian dense flow changes in urban mobility presented in Aranda-Corral et al. (2018) might be easily adapted to BSS by using recorded data. Furthermore, recorded data could be used by Machine Learning solutions for predicting user demand

(near or long term) such as seen in Aranda-Corral et al. (2021). Similarly, USS concept and integration methodology presented in this work could be adapted to other urban mobility fields such as scooter or car share.

Recorded data would be consumed to complement other data science practices on the urban realm (Borrego-Díaz et al., 2014). For example, as an aggregated information system for others that also work with data extracted from third party agents (Miguel-Rodríguez et al., 2016).

On the one hand, stored information could help in providing a prediction on bike availability near to the user as it might be directly published through smartphone apps to improve consumer satisfaction (Froehlich et al., 2009). On the other hand, it is useful to enrich alternatives solutions to unbalanced problems based on economically incentivizing users to return bikes on empty stations as in Singla et al. (2015) would be also supported by recorded data in this work.

Finally, based on recorded data, it would be developed data solutions for global analysis of smart sharing in cities (topology of sharing in urban networks) (Bütter et al., 2011) as well as well-known BSS problem such as unbalanced stations (Brinkmann, 2020). Besides that, it could be faced more ambitious goals, such as the estimation of carbon footprint reduction.

7. Conclusions and future work

The main contribution of his work is to define the "unified station status" (USS), which could establish a bridge between others BSS available storage methodologies found and serves as a merging tool for other BSS data not extracted in this platform, yet. Based on USS, a full BSS information integration platform has been developed.

Despite the main motivation was to provide data to be used by academia in fields related with Artificial Intelligence techniques, in the end, implemented functionality has been extended to a wider range of uses cases. Thus, it covers both side of data accessibility approaches: machine–machine and human-computer communications.

As the developed ETL process has been working since December 2015 (in earlier versions), more than seven years of status records have been stored from near thrifty systems. Consequently, future research common issues, such as making comparison between systems, might have been simplified as its historical information could be retrieved in the same format and meaning.

Concerning the big-data perspective (Jia et al., 2017), an efficient solution for the difficulties intrinsic to the use of BSS data in the ETL process has been developed. Refined data architecture and extraction methodology presented in this study reduces final records in a 93% (from 2.8 billion to 220 million approximately). Moreover, emerging problems of applying this huge data reduction, involving data accessibility and comparison task, has been also overcome through the implemented consumption tools.

Lastly, with respect to ETL phase, BSS information not only has been collected, but also aggregated with external factors, enabling future research to measure its influence on system usage and behaviours patterns, enhancing final results.

Through the API and the web application developed, there is covered the whole data access issue involving different end–users profiles. As a first step, general information about system status is given for end users. Then, deeper information with relevant KPIs related with historical system and station patterns could help in research and maintenance tasks for advanced users from own system manager or operators to scientific or expert focused on the field.

In relation to future work, it would focus on several paths. On the one hand, the authors aim to use stored data and knowledge for feeding Artificial Intelligence models for BSS usage prediction (Machine Learning) and simulation (multi-agent systems). On the other hand, in the medium term, non-AI users are expected to use developed tools for several purposes such as current system availability check or quality of service monitoring.

Finally, it must be mentioned that other BSS standards could be formally related to our proposal to achieve a better interoperability and, hence, to fulfill truly linked open data. In relation to improve the platform itself, new KPIs could be included to developed dashboards. Furthermore, emergent storing methodologies will be taken into consideration to outperform implemented solution with big-data state-of-the-art technologies.

Conflict of interest

The authors do not have any conflict of interest with other entities or researchers.

Acknowledgments

This research was supported by project PID2019-109152 GB-I00 financed by Ministerio de Ciencia, Innovación y Universidades, Spain (MCIN/AEI/10.13039/501100011033), and by project UHU-1266216 (FEDER 2014e2020) financed by Junta de Andalucía and Universidad de Huelva.

REFERENCES

Andrienko, G., Andriendo, N., Boldrini, C., et al., 2021. (So) big data and the transformation of the city. International Journal of Data Science and Analytics 11 (4), 311–340.

Aranda-Corral, G.A., Borrego-Díaz, J., Galán-Páez, J., 2018. Synthetizing qualitative (logical) patterns for pedestrian simulation from data. In: SAI Intelligent Systems Conference (IntelliSys) 2016, London, 2016.

Aranda-Corral, G.A., Rodríguez, M.A., Fernández de Viana, I.A.M.I.G., 2021. Genetic hybrid optimization of a real bike sharing system. Mathematics 9 (18), https://doi.org/10.3390/math9182227.

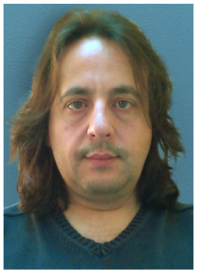
Ashqar, H.I., Elhenawy, M., Rakha, H.A., 2019. Modeling bike counts in a bike-sharing system considering the effect of weather conditions. Case Studies on Transport Policy 7 (2), 261–268.

Barbet, T., Nacer-Weill, A., Yang, C., et al., 2021. An agent-based model for modal shift in public transport. arXiv 2021, https://doi.org/10.48550/arxiv.2107.11399.

Batty, M., 2009. Cities as complex systems: scaling, interaction, networks, dynamics and urban morphologies. In: Meyers, R.A. (Ed.), Encyclopedia of Complexity and Systems Science. Springer, Berlin, pp. 1041–1071.
Bean, R., Pojani, D., Corcoran, J., 2021. How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones. Journal of Transport Geography 95, 103155.
Borrego-Díaz, J., Chávez-González, A.M., Martín-Pérez, M.A., et al., 2012. Semantic geodemography and urban interoperability. In: Research Conference on Metadata and Semantics Research, Berlin, 2012.
Borrego-Díaz, J., Galán-Páez, J., Miguel-Rodríguez, J.D., 2014. Building knowledge layers and networks from urban digital information. In: International Conference Virtual City and Territory 9 Congresso Città e Territorio Virtuale, Roma, 2014.
Brinkmann, J., 2020. Active Balancing of Bike Sharing Systems. Springer, Berlin.
Bütter, J., Mlasowsky, H., Birkholz, T., et al., 2011. Optimising Bike Sharing in European Cities: a Handbook. European Commission, Brussel.
Caldarelli, G., Arcaute, E., Barthelemy, M., et al., 2023. The role of complexity for digital twins of cities. Nature Computational Science 3, 374–381.
Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., et al., 2018. Open data interoperability. In: Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., et al. (Eds.), The World of Open Data. Springer, Berlin, pp. 75–93.
Chemla, D., Meunier, F., Wolfler Calvo, R., 2013. Bike sharing systems: solving the static rebalancing problem. Discrete Optimization 10 (2), 120–146.
Costa, E., Silvestri, F., 2021. On the bike spreading problem. In: 21st Symposium on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems (ATMOS 2021), Dagstnhl, 2021.
Daraei, S., Pelechrinis, K., Quercia, D., 2021. A data-driven approach for assessing biking safety in cities. EPJ Data Science 10 (1), 11.
Dell'Amico, M., Hadjicostantinou, E., Iori, M., et al., 2014. The bike sharing rebalancing problem: mathematical formulations and benchmark instances. Omega 45, 7–19.
Erdogan, G., Laporte, G., Calvo, R.W., 2014. The static bicycle relocation problem with demand intervals. European Journal of Operational Research 238 (2), 451–457.
Federico, C., Chiara, P., Andrea, Z., et al., 2018. A dynamic approach to rebalancing bike-sharing systems. Sensors 18 (2), 512–533.
Ferenchak, N.N., 2023. Longitudinal bicyclist, driver, and pedestrian perceptions of autonomous vehicle communication strategies. Journal of Traffic and Transportation Engineering (English Edition) 10 (1), 31–44.
Froehlich, J., Neumann, J., Oliver, N., 2009. Sensing and predicting the pulse of the city through shared bicycling. In: 21st International Joint Conference on Artificial Intelligence, IJCAI'09, Pasadena, 2009.
Jia, Z., Xie, G., Gao, J., et al., 2017. Bike-sharing system: a big-data perspective. In: Qiu, M. (Ed.), Smart Computing and Communication. Springer, Berlin, pp. 548–557.
Jiménez, P., Nogal, M., 2021. Analysis of real experiences using different sized bike sharing schemes in Irish cities. In: XIV Conference on Transport Engineering (CIT2021), Burgos, 2021.
Jurdak, R., 2013. The impact of cost and network topology on urban mobility: a study of public bicycle usage in 2 U.S. cities. Plos One 8 (11), 1–6.
Li, Q., Fan, R., Qian, Z., 2017. A nonlinear solution to closed queueing networks for bike sharing systems with Markovian arrival processes and under an irreducible path graph. In: Yue, W., Li, Q., Jin, S., et al. (Eds.), Queueing Theory and Network Applications. Springer, Berlin, pp. 118–140.
Li, X., Gong, Y., Liu, W., et al., 2024. Dual-track spatio-temporal learning for urban flow prediction with adaptive normalization. Artificial Intelligence 328, 104065.
Li, J., Xu, C., 2024. Evidence-based practices in sustainable travel behavior intervention: a knowledge graph-based systematic review. Journal of Traffic and Transportation Engineering (English Edition) 11 (2), 293–311.
Lim, H., Chung, K., Lee, S., 2022. Probabilistic forecasting for demand of a bike-sharing service using a deep-learning approach. Sustainability 14 (23), 15889.
Liu, C., Gao, X., Wang, X., 2022a. Data adaptive functional outlier detection: analysis of the paris bike sharing system data. Information Sciences 602, 13–42.
Liu, S., Shen, Z., Ji, X., 2022b. Urban bike lane planning with bike trajectories: models, algorithms, and a real-world case study. Manufacturing & Service Operations Management 24 (5), 2500–2515.
Makarova, I., Shubenkova, K., Pashkevich, A., et al., 2017. Smartbike as one of the ways to ensure sustainable mobility in smart cities. In: Magno, M., Ferrero, F., Bilas, V. (Eds.), Sensor Systems and Software. Springer, Berlin, pp. 181–198.
Marazzini, M., Mitolo, N., Nesi, P., et al., 2018. Smart city control room dashboards: big data infrastructure, from data to decision support. Journal of Visual Language and Sentient System 4, 75–82.
Marquez-Saldaña, F.J., Aranda-Corral, G.A., Borrego-Díaz, J., 2022. Enabling knowledge extraction on bike sharing systems throughout open data. In: HCI in Mobility, Transport, and Automotive Systems: 4th International Conference (MobiTAS 2022), Berlin, 2022.
Miguel-Rodríguez, J.D., Galán-Páez, J., Aranda-Corral, G.A., et al., 2016. Urban knowledge extraction, representation and reasoning as a bridge from data city towards smart city. In: 2016 International IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing. Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulonse, 2016.
Millett, L.I., Estrin, D.L., et al., 2012. Computing Research for Sustainability. The National Academies Press, Washington DC.
Morton, S., Pencheon, D., Squires, N., 2017. Sustainable development goals (SDGs), and their implementation: a national global framework for health, development and equity needs a systems approach at every level. British Medical Bulletin 124 (1), 81–90.
Natera Orozco, L.G., Battiston, F., Iñiguez, G., et al., 2020. Data-driven strategies for optimal bicycle network growth. Royal Society Open Science 7 (12), 201130.
Pearson, L., Dipnall, J., Gabbe, B., et al., 2022. The potential for bike riding across entire cities: quantifying spatial variation in interest in bike riding. Journal of Transport & Health 24, 101290.
Reggiani, G., Van Oijen, T., Hamedmoghadam, H., et al., 2022. Understanding bikeability: a methodology to assess urban networks. Transportation 49 (3), 897–925.
Rennie, N., Cleophas, C., Sykulski, A.M., et al., 2022. Analysing and visualising bike sharing demand with outliers. ArXiv 2204, 06112.
Ricci, M., 2015. Bike sharing: a review of evidence on impacts and processes of implementation and operation. Research in Transportation Business & Management 15, 28–38.
Sanmiguel-Rodríguez, A., Giráldez, V.A., 2019. Impact of climate on a bike-sharing system. minutes of use depending on day of the week, month and season of the year. Cuadernos de Psicología del Deporte 19 (2), 102–112.
Shimizu, S., Akai, K., Nishino, N., 2013. Modeling and multi-agent simulation of bicycle sharing. In: 1st International Conference of Serviceology (ICServ) 2013, Tokyo, 2013.

Singla, A., Santoni, M., Bartók, G., et al., 2015. Incentivizing users for balancing bike sharing systems. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15), Austin, 2015.
Szell, M., Mimar, S., Perlman, T., et al., 2021. Growing urban bicycle networks. ArXiv, 2107.02185V2.
Valle, E.D., Celino, I., Dell'Aglio, D., 2010. The experience of realizing a semantic web urban computing application. Transactions in GIS 14 (2), 163–181.
Vogel, P., 2016. Service Network Design of Bike Sharing Systems: Analysis and Optimization. Lecture Notes in Mobility. Springer, Berlin.
Yang, C., Raskin, R., Goodchild, M., et al., 2010. Geospatial cyberinfrastructure: past, present and future. Computers, Environment and Urban Systems 34 (4), 264–277.
Yaqoob, S., Cafiso, S., Morabito, G., et al., 2023. Detection of anomalies in cycling behavior with convolutional neural network and deep learning. European Transport Research Review 15 (1), 9.
Zhou, S., Song, C., Wang, T., et al., 2022. A short-term hybrid TCN-GRU prediction model of bike-sharing demand based on travel characteristics mining. Entropy 24 (9), 1193.

Francisco Márquez-Saldaña holds a degree in computer science and artificial intelligence from Huelva University. He is a PhD candidate studying at University of Seville. He mainly engaged in research on data analytics, machine learning applications and multi-agent system simulations in bike sharing systems. Moreover, he has participated in a research project related to agent-based modelling in urban systems.



Dr. Gonzalo A. Aranda-Corral holds a degree in physics from Universidad Complutense de Madrid, a master's degree in advanced computer technologies from the University of Huelva, and a PhD in logic, computation and artificial intelligence from the University of Seville, awarded Cum Laude. His research, conducted within the "logic, computation and knowledge engineering" group, focuses on multi-agent systems and distributed knowledge processing, with applications in the semantic web and intelligent urban planning. He has over 3500 h of university teaching experience in computer science and artificial intelligence, emphasizing the use of ICT for enhanced student interaction.

Dr. Joaquín Borrego-Díaz is the director of the LOCIC research group (PAIDI TIC-137: logic, computation, and knowledge engineering). His research focuses on the intersection of logic, computation, and artificial intelligence, specifically in the areas of the theory of computation, models of computation, computability, and knowledge representation and reasoning as classified by the ACM computing classification system. His work stands out within the field of computer science and artificial intelligence (CCIA). Notable contributions include exploring formal ontologies, formal concept analysis (FCA), automated theorem proving, verification, multi-agent systems, and developing a learning theory based on FCA.