Project co-financed by the European Regional Development Fund

Deliverable 3.2.1: *Proposal for a standard system of Big Data sets available at Med level*

Date: 30/11/2020

WP3 STUDYING

Activity 3.2

"Establishing common standards and tools for observatories"

Requested by:

REGIONE LAZIO

Developed by:

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
CENTER FOR ADVANCED STUDIES IN TOURISM

Research Authors:

This report has been developed by CAST research team (Cristina Bernini, Maria Laura Gasparini, Alessia Mariotti, Valeria Villalobos) as external experts of Regione Lazio project partner.

BEST MED project partners:

Lead partner: *El legado andalusí Andalusian Public Foundation*

Project co-financed by the European Regional Development Fund

Table of Contents

Interreg Mediterranean BEST MED

Project co-financed by the European Regional Development Fund

List of Tables

List of Figures

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Introduction

BEST MED, *Beyond European Sustainable Tourism MED Path* is being implemented in eight Mediterranean countries (Spain, Portugal, France, Italy, Croatia, Slovenia, Greece and Montenegro) with the general objective of enhancing Mediterranean Governance, being the main challenges to fight against seasonality and lack of effective cooperation among main tourism actors, including the citizen active participation on the policies design. It aims to have a new integrated and sustainable tourism planning approach, to contribute to the mitigation of seasonality in the MED area, through the connection between coastal and inland regions, such as a path-route method. A testing phase will allow to build a joint model that will be transferred and capitalised, as well as a toolkit and updates set of data indicators.

BEST MED will follow a strategy of previous approaches and outputs, testing an updated toolkit of data and indicators, contributing to the design of a new Green model (MED S&C Path- Sustainable Path & Cultural Routes Model), focusing on integration of tourism planning into wider development strategies, together with mobilizing key players both at local and specifically at transnational level, creating synergies across MED countries and promoting the awareness of the MED area. More information about the project can be found here.

The objective of the **Working Package 3** (WP3) is to develop a knowledge framework related to the main project goals through:

- Base information for a network of tourism observatories.
- Information needed to develop a MED Sustainable Path and Cultural Routes Model (MED S&C Path) on the example of the Mitomed+ project "Green Beach Model", and other MED projects.

The study will examine existing methodological approaches on tourism data and tourism observatories and analyse previous experiences on tourism data knowledge, finding gaps and needs in data collection management and pinpointing the main results and suggestions from the previous MED projects, to develop adequate policies.

Within WP3, the activity 3.2 *"Establishing common standards and tools for observatories"* aims to establish base information for developing institutional

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

background of data management as a Mediterranean Network of national and regional Tourism Observatories, expanding cooperation of existing observatories for developing institutional background of data management.

To this end, the **Deliverable 3.2.1: Definition of a thematic data collection system** is divided into two key sections:

- **Big data**: proposal for a standard system of Big Data sets available at Med level.
- **Green Path destination indicators**: standard set of indicators based on previous experience of MITOMED+ "Green Beach Model", adapted to the "MED S&C Path Model".

The present report contains the first section of Deliverable 3.2.1, and thus it focuses on the possible uses of big data in the tourism sector at MED level. As such, the report is organized as follows:

- Section 1: Big Data.
- Section 2: Big Data Initiatives in Tourism Relevant for the MED .
- Section 3: Recommendations and Conclusions.

First, section 1 establishes, in general terms, the theoretical framework related to the use of big data sources in tourism. In this way, the potential types of tourism big data are defined and illustrated through relevant examples in the areas of sustainable tourism, official tourism statistics, tourist mobility, accommodation, among many others. Special attention is given to Eurostat's ESSnet Big Data II project, and in particular its Working Package J, *Innovative Tourism Statistics*, since it aims to resolve the need for a conceptual framework in tourism big data by creating an EU-wide tourism information system.

Next, section 2 provides an overview of the main aspects of numerous big data initiatives in tourism at MED level. More in-depth information about these particular initiatives, as well as other interesting tourism big data projects outside the MED region can be found in the accompanying Excel file for this report.

Finally, section 3 outlines the most relevant recommendations on the use of new data sources in the tourism sector, structured according to the type of big data source, and the feasibility of their implementation.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

1 Big Data

As a result of accelerated advances in computer science and Internet technology, a huge array of unstructured and structured data are currently being generated and processed, contributing to the arrival of the Big Data era. Big Data can be understood as a discipline devoted to the study, distribution, and preservation of vast data sets that often derive from a variety of sources. However, despite the implementation and common use of big data in different industries, the actual definition is still a controversial topic. The most commonly recognised concept of big data includes its three core features, generally referred to as the "3 Vs" of big data: *volume*, *velocity*, and *variety* (J. Li et al., 2018). Other terms later attributed to the concept of big data are *veracity* and *value*, summarised by Erl et al. (2016) as follows:

a) **Volume**: The amount of data processed by Big Data systems is significant and ever-increasing. High data volumes place specific data storage and processing requirements, as well as additional data preparation and management procedures. Common data sources responsible for producing high volumes of data may include, among others, electronic payment transactions, such as point-of-sale and banking; science and research projects such as the Large Hadron Collider; sensors, including GPS sensors, radio-frequency identification devices, and smart metres; and social media platforms like Facebook and Twitter.

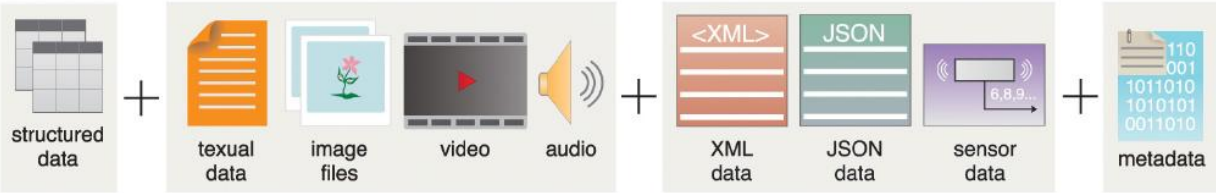
b) **Velocity**: Data can arrive at incredible high rates of speed in big data environments, thus massive datasets can be accumulated in little time. In this case, managing the fast inflow of data requires developing highly adaptable and accessible data processing systems and storage capabilities. To further illustrate the notion of velocity, the following data can be quickly produced in any given minute: 350,000 tweets, 300 hours of video footage posted to YouTube, 171 million emails, and 330 GB of jet engine sensor data.

c) **Variety**: Variety refers to the various formats and data types that need to be supported by Big Data solutions. For instance, as shown in Figure 1, data variety can incorporate structured data, semi-structured data (XML, JSON, and sensor data), unstructured data (text, image, video, and audio files), and their corresponding metadata. Still, data variety can generate data integration,

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

transformation, processing, and storage problems that need to be carefully addressed.

Figure 1. Example of variety in big data



Note. Reprinted from Big data fundamentals: concepts, drivers & techniques (p. 31), by T. Erl et al., 2016. Prentice Hall.

d) **Veracity**: Veracity refers to the consistency or quality of the data. In this regard, data has to be tested for accuracy, which can lead to data processing activities intended to address invalid data and removing noise. In terms of veracity, data can be designated as either signal or noise. Noise is data that cannot be translated into information and thus has no relevance, while signals have the potential for contributing to useful information. Data with a high signal-to-noise ratio are more reliable than data with a lower ratio. Data obtained in a standardized way typically produce less noise than data gathered through unregulated sources, such as social media posts.

e) **Value**: Value is characterised as the usefulness of data; therefore, the value characteristic is logically connected to the veracity characteristic mentioned above, since the higher the data accuracy, the higher the value. Similarly, value also refers to the amount of time it takes to accurately analyse the data, and other lifecycle concerns such as how well the data were stored, if any useful qualities were lost during data cleaning, if the correct questions were asked during data analysis, and if the findings were corrected transmitted to the relevant decision-makers.

All things considered, the benefits of these modern data sources over conventional methods of knowledge generation in tourism also lie in these characteristics, because tourism big data are typically comprehensive (large *volume*), easily accessible (*velocity*) and range from a multitude of different sources (structured, semi-structured and unstructured, *variety*).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

In addition, an equally important aspect is the different typologies of available big data. First, data can be either *machine-generated*, or *human-generated*. Machine-generated data is the largest source of big data, and it refers to data produced directly by software programs or hardware devices such as sensors, satellites, log files, bioinformatics, trackers, personal health tracker and many other meaningful data tools (Ghotkar & Rokde, 2016). On the other hand, human-generated data is the product of direct human interactions with technologies such as digital services and devices, including social media, emails, and messages.

As a result, human-generated and machined-generated data can be either *structured*, *unstructured*, or *semi-structured*. In essence, structured data refers to data that can be easily collected, stored, and analysed. Therefore, structured data, such as financial transfers, invoices, and customer information, is often stored in tabular form in a relational database. Instead, unstructured data refers to data lacking a recognisable form, and it is what comprises the majority of available big data (videos, images, audio, etc.) (Syed et al., 2013). Finally, semi-structured data has a given structure and consistency, but is not relational in nature, but rather hierarchical or graphical. In this case, such data is usually stored as JSON or XML files.

Similarly, *metadata* is essential for big data collection, storage, and interpretation as it provides information about the quality and origin of the data, especially when processing semi-structured or unstructured data. Metadata includes the attributes and structure of a dataset, for example, XML tags containing author and document formation date, and attributes concerning the file size and resolution of a picture (Erl et al., 2016).

Big data analysis integrates conventional statistical approaches with computer science methods. Statistical sampling from a population is appropriate when the entire dataset is available, which is characteristic of conventional “batch” data processing scenarios. In contrast, big data allows to process data in real time through different types of data analysis such as data mining, statistical analysis (A/B testing, correlation, regression), machine learning (classification, clustering, outlier detection, filtering, deep leaning), semantic analysis (natural language processing, text analytics, sentiment analysis), and visual analysis (heat maps, time series plots, network graphs, spatial data mapping) (Erl et al., 2016).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Some of the most relevant methods for tourism are briefly explained below, as stated by Feldman & Sanger (2007):

a) **Machine learning**: Traditionally, a machine learning workflow consists of different aspects. First, data is collected, prepared, and stored as "training", "testing" and "validation" datasets, then a model is trained (or "learns" from the data), and finally the model is tested and updated accordingly. The workflow is also assumed to be recursive since the forecasts inform the description of the problem and the simulation techniques used, as well as the historical evidence itself, will change with time. Therefore, the purpose of continuous feedback within the machine learning process is to increase accuracy.

b) **Semantic analysis**: Semantic analysis refers to the process of extracting relevant information from textual data. In a practical sense, it aims to derive valuable knowledge from data sources by finding and investigating important trends. However, in the case of semantic analysis these data sources are text collections, and significant patterns are found not in structured database records, but in unstructured textual data.

Overall, given the complexity of big data coupled with its powerful analytical capability, there are several problems that need to be understood and prepared before implementing big data solutions. For instance, issues related to data protection, security, anonymity, and quality need to be carefully planned. In fact, big data also opens up new ways to explore going beyond on-site systems to remotely provided, flexible environments based in the cloud (Erl et al., 2016). Indeed, all of the above aspects require any organisation to identify and develop a collection of governance mechanisms and decision structures to ensure that appropriate actors understand the existence, consequences, and specifications of big data.

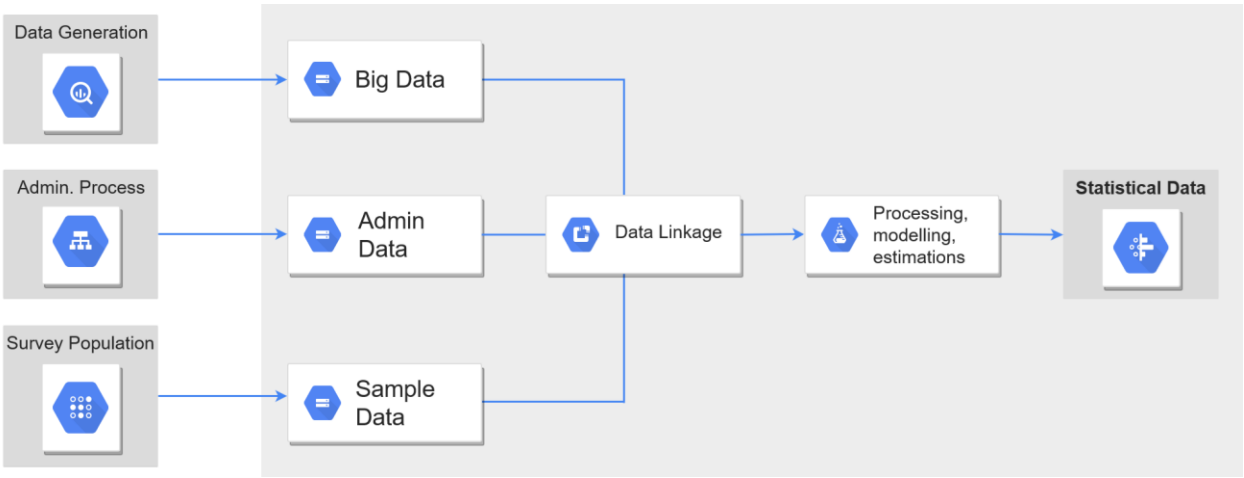
1.1 Integration of Big Data in Official Statistics

Official statistics include four different data sources: survey data, census data, administrative data, and more recently, big data. In the case of official statistics, big data can be used as auxiliary data, combined with other statistical data from surveys or from institutional sources. Two different instances can be distinguished: the big data source can be linked to statistical data at the individual-level, or at any intermediate level of aggregation, such as country or sector level (Braaksma & Zeelenberg, 2020).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Most NSIs work with big data and other data sources, particularly administrative data, to supplement surveys. One predominant theme is the need to merge data sources such as polls and censuses, administrative data, and large and not-so-big data, including medical record systems, commercially compiled data, financial data, transaction data, handheld devices, wearable measuring devices, web scraping, blogs, social media, smart metres, satellites, global positioning system (GPS) data, sensors, pictures and videos, vessel and flight data, and scanner data (Japec & Lyberg, 2020). An example of a general framework for statistical production in a multi-source context is depicted in Figure 2.

Figure 2. Example of a framework for the integration of big data in official statistics



Note. Adapted from "Use of Big Data in official statistics" by G. Barcaroli, 2015, *ISI World Statistics Congress*, p. 3.

However, there is still no definitive, widely agreed framework for assessing the quality of big data for its integration in official statistics. In this regard, Braaksma et al. (2020) note that some work has already been done on improving international big data quality systems, in particular the ground-breaking work done in the UN-ECE Big Data Quality Task Team, the American Association of Public Opinion Research (AAPOR) Big Data Task Force, and Eurostat's ESSnet Big Data projects. An overview of the main policy challenges for integrating big data into official statistics can be found in Table 1.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Table 1. Big data challenges in official statistics

Type	Challenge	Description
Policy Challenge	**Data Ownership**	Many users already create data that is theoretically valuable for analysis as part of their day-to-day digital presence. But there has often been a lack of consistency in legal guidance due to a lack of clarity as to who controls the data. These new forms of data make ownership laws much more unclear: data are no longer housed in statistical organisations, with well-defined rules of conduct, but are housed in corporations or administrative agencies. Therefore, it will be essential for NSIs to remain updated about evolving laws and to be mindful of variations in legislation across countries.
Policy Challenge	**Data Stewardship**	In the past, the emphasis has been almost entirely on improving methodologies to enhance the predictive use of survey data and, to a lesser degree, administrative data. It is important to extend training efforts to prepare scientists to increase awareness of issues such as identifying the relevant population and methodologies for data linkage. However, it is important to incorporate the teaching of these skills into current programmes, particularly as the field moves towards data integration between survey and non-survey data.
Policy Challenge	**Data Collection Authority**	When statistical agencies were the major data collectors, they did so with quite simple contractual, statutory protections. The regulatory data collection authorization for the new technologies is less evident. Thus, there are important concerns about what is fairly private and what constitutes unjustified interference.
Policy Challenge	**Privacy and Re-identification**	The risk of re-identification has grown due to increasing public access of identified data and rapid developments in data-linking technologies. Since big data has broad scope, it is not possible to rely on rely on protection from sampling. For example, a person with knowledge of the zip code, birthdate and gender of an individual may re-identify more than 80% of Netflix users, but none of that information is usually classified as Personally Identifiable Information (PII).
Technical Challenge	**Skills for Big Data Integration**	Depending on the size of the data being examined, there could be major obstacles in terms of the expertise and services available to deal with big data. Usually big data projects require at least four different roles: a domain expert, a researcher, a computer scientist, and a system administrator.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Technology Challenge	**Computational Requirements**	Organizations looking to experiment with big data computing cluster technologies will minimise their initial capital outlays by renting pre-built computed cluster infrastructure (such as Apache Hadoop) from online vendors such as Amazon Web Services. Systems such as Apache Hadoop significantly simplify the development of computing clusters capable of facilitating parallel processing of big data computations. While the cost of magnetic storage media may be low, the cost of designing solutions for long-term storage and processing of big data remains high. The use of external cluster services is a short-term solution to this problem.

Note. Adapted from "AAPOR Report on Big Data" by AAPOR, 2015, pp. 24-30.

As a matter of fact, big data can be extremely unpredictable and selective: their population coverage can change from day to day, contributing to inexplicable time-series changes. Also, most of the times individual observations in these big data sets do not consider linking variables and therefore cannot be easily linked to other datasets. In other words, for big data, there is often inadequate knowledge about the data source, but in this case, certain models such as multiple regression models or logit models, are then helpful in formulating clear conclusions about these relationships and calculating selectivity or coverage characteristics.

Another approach is to use more advanced methods for big data, including high-dimensional regression, machine learning techniques, graphical modelling, data science, and Bayesian networks. Or even more traditional methods, such as Bayesian analytics, time-series methods and multilevel models, and time-series models. A third strategy is to take draw inspiration from the method in which national accounts are generally collected. In this context, many sources that are usually missing, imperfect and/or partially conflicting are combined, using a statistical reference frame to gain a holistic view of the whole economy. Similarly, big data and other sources that are potentially unreliable or skewed can be merged to provide a complete and accurate representation of a given phenomenon (Braaksma et al., 2020).

For example, according to Tam & Van Halderen (2020), over the past decade, research has found several opportunities for the use of transaction data, otherwise known as scanner data, in compiling the Consumer Price Index (CPI). In addition, new approaches, known as multilateral index methods (including the GEKS method, and the Geary-

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Khamis method[1], among others), have been developed to compile the CPI using transaction data. The consensus is that multilateral index approaches are the most successful way to leverage the maximum amount of information available in the transaction records. Similarly, satellite imagery data or Earth Observation (EO) data, can be used in many areas of official statistics. Specifically, in agricultural statistics for calculating land cover and land use, crop classification and crop yields, sustainable development indicators, and more. And finally, Mobile Network Operators (MNO) data have the largest potential to generate insights into official statistics. MNO data can be used to collect tourist statistics, demographic statistics, migration statistics, commuting statistics, traffic flow statistics, and data related to border employment and seasonal workers.

Daas et al. (2015) affirm that official statistics will benefit tremendously from the opportunities presented by big data. However, caution is required when attempting to translate these sources into official figures, and potential issues such as missing data, volatility and selectivity need to be properly addressed. For this purpose, NSIs need to engage in relevant research and continuously develop and update big data skills. In addition, big data teams should be multi-disciplinary and supported by experts in the fields of data mining, big data processing, high-performance computing, and data science. Thus, when created in a methodologically sound way, official big data statistics can be cheaper, quicker, and more accurate than the official statistics known to date. But in order for these efforts to be successful, it is important that they are embraced by the general public and by local and European legislation.

In short, despite the absence of a clearly specified framework for the integration of big data into official statistics, significant steps are being taken at the European level. A future paradigm shift will then have to deal with the incorporation of big data in time and space, develop new methods for data visualisation, take into account that indicators are measured differently, and that various data quality concerns are at stake. It would also require careful supervision of the protection of privacy measures in the process (Braaksma et al., 2020).

[1] A more in-depth discussion of multilateral index methods can be found in Chessa et al. (2017).

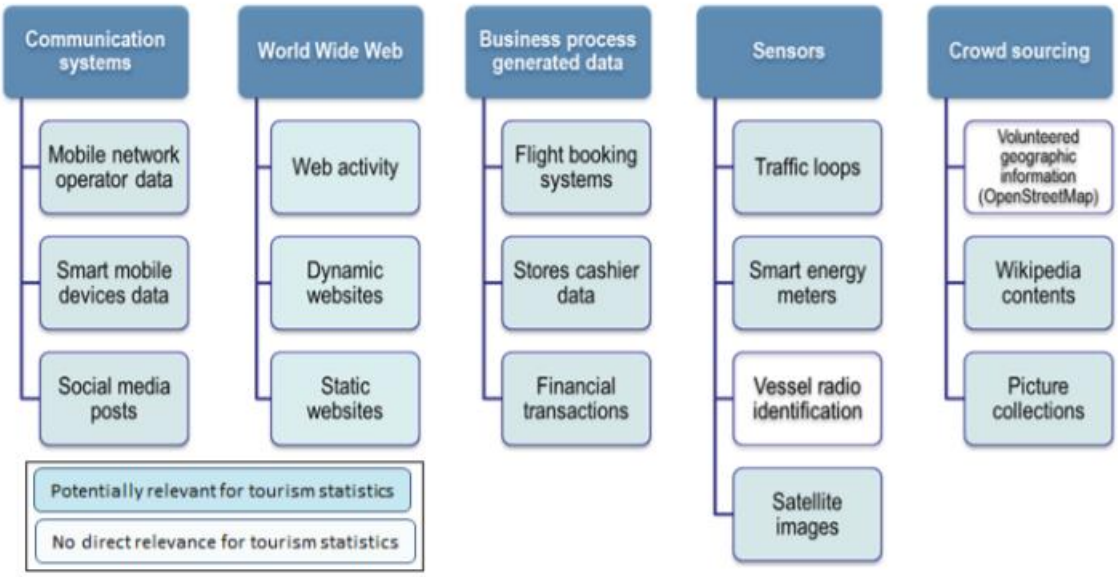
Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

1.2 Big Data in Tourism

1.2.1 Sources

Within the context of tourism, big data are considered a secondary source, as they are often generated for a technical purpose other than tourism, such as mobile data or social media information, for example. In this sense, big data can be collected and analysed for tourism research purposes (Reif & Schmücker, 2020). As shown in Figure 3, big data sources for tourism can therefore be classified into five areas, depending on how the data are generated: communication systems, world wide web, business process generated data, sensors, and crowd sourcing (Demunter, 2017).

Figure 3. Taxonomy of big data sources relevant for tourism



Note. Reprinted from Tourism statistics: Early adopters of big data? (p. 9), by C. Demunter, 2017. Publications Office of the European Union.

First, big data sources concerning communication systems are related to mobile positioning data produced by mobile network operators and intelligent mobile devices, as well as information generated from social media posts. The second group focuses on the internet as a source of big data, including information from search engines and page views (for instance Google Trends data and Wikipedia page views), dynamic websites (TripAdvisor, Booking.com, and Airbnb, for example), and static websites

Project co-financed by the European Regional Development Fund

which include constant information such as the location of tourist establishments, the number of rooms, etc.

Business process generated data refers to the constant stream of data produced through regular business processes, in particular, flight booking systems such as Amadeus can be used to complement tourism demand surveys; store cashiers data can be implemented as a proxy for tourism seasonality considering the variations in turnover; and sources of financial transactions involve using payment card data to measure monetary information related to tourism, however, this particular type of data source has yet to be widely implemented due to security concerns.

Another important source of big data is related to the use of sensors to monitor people's movements, land use, energy consumption, and so forth. Given the characteristics of this specific type of data source, sensor-generated data has a significant potential for measuring sustainable tourism indicators. And finally, crowdsourcing refers to user-generated content relevant for tourism statistics. For example, geo-location data available in images published online by tourists, and page views and geo-location data from Wikipedia articles on tourist destinations.

Currently, tourism statistics are mostly compiled by using surveys as the main data sources, as advised by the UNWTO's international recommendations for tourism statistics, and Eurostat's methodological manual for tourism statistics. However, the new sources of big data described above represent a possibility to improve and enrich the existing system of tourism statistics. And as such, the major goal is to transform the tourism statistics system into a data factory that uses a wide range of input sources to serve several output needs simultaneously (Demunter, 2017).

1.2.2 State of the Art

Information and knowledge are indispensable for the progress of strategic and innovative initiatives, which are necessary for decision-making and productivity. Big data, like any innovation process, entails a range of intangible gains for organisations, such as expanding the knowledge base; enhancing human resource skills; and pursuing the development of new opportunities. The introduction of big data technologies to an organisation requires a comprehensive review of all forms of data, in order to define transformation techniques, to model on real-time or near-real-time decision-making, and to incorporate new appraisal parameters (González Gómez & Rubio Gil, 2020). Thus, in

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

order to properly evaluate the potential of big data within the scope of the BEST MED project, it is necessary to assess the current state of tourism research with respect to this particular subject.

In recent years there has been considerable interest in the study of big data in tourism research. Accordingly, several literature reviews have been carried out to evaluate the current research scenario pertaining to the use of big data in tourism. For example, González Gómez & Rubio Gil (2020) carried out a bibliometric analysis of the Scopus, Web of Science and Science Direct databases for the period 2015-2019, considering keywords such as big data, tourism, knowledge, and value. The purpose of this analysis was to determine the current state of tourism research on the use of big data, with a particular emphasis on its added value and knowledge generation for the tourism sector. As a result, four major research categories were identified:

a) **Tourist destinations**: Evaluation of the tourist destination image based on the analysis of social media data, unstructured data (images and videos), MNO data, and assessments of web pages.

b) **Consumer analytics**: Analysis of data aimed at explaining the profile of the customer, that is, the analysis of the tourist's behaviour, feelings, and experiences.
c) **Technologies**: Application of big data tools in the tourism sector, and the development of new methodologies for the use and analysis of big data.

d) **Others**: The final category, called "Others", comprises the analysis of tourism organizations, as well as big data challenges and opportunities.

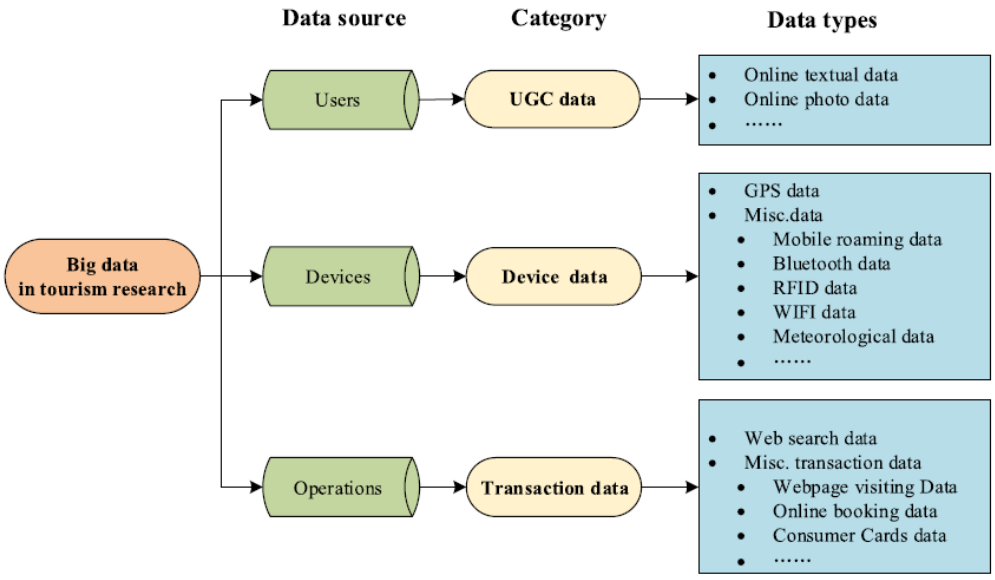
In this case, the most important contribution by far is related to the analysis of social media and booking platforms data, as a result of the automation and evolution of the tourism intermediation sector (travel agencies and tour operators).

Similarly, Li et al. (2018) conducted a comprehensive literature review with a particular focus on the different types of big data predominantly used in tourism research. In this sense, as illustrated in Figure 4, the authors identified three major categories of big data sources in tourism: *user-generated content (UGC) data*, including text information and photo data publicly available online; *device data*, in particular GPS data (by device),

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

roaming data, and Bluetooth data; and *transaction data*, specifically web search results, data on website visits, and data on online bookings.

Figure 4. Categories of big data sources in tourism research



Note. Reprinted from "Big data in tourism research: A literature review" by J. Li et al., 2018, *Tourism Management*, 68, p. 305. (https://doi.org/10.1016/j.tourman.2018.03.009)

Interestingly, the authors also point out the evident rapid evolution of the use of big data in tourism. As a matter of fact, a limited number of studies began to emerge during 2007, demonstrating a steady pattern of growth over the next few years, and peaking during 2015 and 2016. Furthermore, Asia, Europe and North America have made the biggest contribution to the application of big data to tourism science, specifically, USA ranked first in terms of number of publications, followed by China and Australia. As far as European countries are concerned, important contributions have been made by Italy, Spain, Belgium, Austria, and The Netherlands.

In their analysis, the authors state that the dominant category is UGC data, which has greatly supported the study of tourist sentiment analysis, tourist behaviour analysis, tourism marketing, and tourism recommendation. Followed by device data, which is still at a developing stage, but with promising potential uses for the study of spatial-temporal tourist behaviour. Finally, tourism research using transaction data has been relatively unexplored, the primary explanation lies in the difficulty of obtaining such data, which is largely under the private jurisdiction of tourism organizations or protected by

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

public institutions. Table 2 offers a summarised comparison among the different uses of big data in tourism research.

Table 2. Comparison of the main types of big data used in tourism research

Source	Data Type	Research Focus	Advantages	Disadvantages
Users	**Online textual data**	Tourist sentiment analysis Tourism recommendation Tourist behaviour analysis Tourism marketing Tourism recommendation	Low cost Conveying tourist sentiment	Reliability concerns
	Online photo data		Low cost Containing multi-metadata	Low location precision
Devices	**GPS**	Data feasibility in tourism Tourist spatial-temporal behaviour analysis Tourism recommendation	Global High precision	High cost Privacy concerns
	MNO		Availability in less visited places Allowing tracking tourist origins	Privacy concerns Low precision of location
	Bluetooth		Availability in crowed indoor spaces Allows unannounced tracking	Small-range coverage Privacy concerns
	RFID		Low cost High precision Availability in crowed indoor spaces	Small coverage range Privacy concerns
	Wi-Fi		Availability in crowed indoor spaces Allows unannounced tracking	Small coverage range Privacy concerns
	Meteorological	Effect estimation of weather on tourism Tourism recommendation	Contains weather factors	Difficulty in processing multi-format data
Operations	**Web search**	Tourism demand prediction Search engine optimization	Low cost Reflecting public attention	Possible estimation biases

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

	Other transaction data	Tourist behaviour analysis Tourism marketing	Records tourists' operations in tourism markets	Privacy concerns

Note. Adapted from "Big data in tourism research: A literature review" by J. Li et al., 2018, *Tourism Management*, 68, p. 319. (https://doi.org/10.1016/j.tourman.2018.03.009)

According to Li et al. (2018), even with impressive developments of big data research in tourism, there is still an immense amount of unexplored potential, especially from the perspective of expanding research and improving analytical techniques. In addition, to increase the use of information related to device and transaction data, it is imperative to strengthen the partnerships between tourism researchers, tourism enterprises, and the public sector. Finally, other valuable types of big data, such as audio data, video data, cross-domain data, multi-type data, etc., have not yet been fully explored in tourism research.

Furthermore, Li & Law (2020) offered a detailed network analysis to explain the current state of tourism big data research by investigating multidisciplinary contributions related to big data. A systematic network analytical method, including co-citation, clustering, and pattern analysis, was used to evaluate publications from the period 2008-2017. As the authors explain, in general, big data research focuses on the development of algorithms, analytical tools, and varied applications. Instead, tourism research regarding big data is largely concerned with application-based studies (such as social media analysis), and only a few theory-based studies.

Specifically, as explained by Li & Law (2020), researchers have used econometrics, data mining, and business intelligence techniques to address big data issues in tourism and hospitality, since these methodologies provide reliable forecasts for tourist demand and customer experiences. However, there are only a small number of prevalent theories about using big data in tourism and hospitality. For example, even though many studies have applied online reviews data in an effort to determine the effect of big data on hotel performance, only a few studies have illustrated the role of big data on hotel performance from a theoretical perspective. Therefore, the theoretical foundations of big data regarding tourism and hospitality need to be further strengthened.

Centobelli & Ndou (2019) carried out a systematic literature review to analyse the evolution of scientific research on big data and analytics in tourism. As expected, their findings indicate that there has been a growing pattern in the literature on big data for tourism in recent years. "Business, Management and Accounting", "Social Sciences"

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

and “Computer Science” are the subject areas most researched and potentially also the most interesting for future applications or analysis. In this regard, China is an exemplary country in the area of big data analysis in the tourism industry. This may be explained by the massive amount of data currently produced by Chinese companies, and the large number of online users producing data on a daily basis. Additionally, Italy and the USA both make major contributions to big data research. As for Italy, the Italian big data industry has expanded steadily in recent years and major investments are planned in the future for businesses operating in big data, analytics, business intelligence and data science. Furthermore, as far as Italian scholarly research is concerned, many projects are directed at enhancing people's lives by designing mobility applications, estimating large-scale events, and smart cities initiatives in general.

As the authors point out, it is important to take into consideration not just the technical aspects of big data, but also the organizational characteristics (for example, data-driven culture) and competitive advantage potential. Certainly, working with big data should involve a data-driven community, and modern computational approaches as well as new competences and skills. As a result, three separate thematic areas concerning tourism research on big data were identified:

a) **Impact of big data analytics on tourism business, management, and innovation practices**: Comprising research related to big data for smart tourism destinations, and social media analytics for planning, management, and value creation.

b) **Role of big data for customer knowledge management and performance**: Focusing on consumer-generated data such as opinions, online reviews, blogs, and tagged photos. In this case, research focuses on grasping customer patterns and behaviours, and supporting tourism forecasting through big data.

c) **Technical, methodological, and architectural solutions supporting big data analytics**: Including analytical techniques and solutions, and big data knowledge infrastructures.

In addition, Samara et al. (2020) performed a systematic literature review concerning the advantages and overall role of big data and artificial intelligence (AI) in the tourism sector. The findings show that structured and unstructured data are used to forecast tourist demand, customer satisfaction, attraction ratings, company success, and

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

revenue. Moreover, automation, robotics, augmented personnel, and digitalized-augmented processes are employed to accelerate daily tasks. However, the implementation of big data and AI technologies in the tourism sector remains low, as it is hindered by several challenges. These challenges may be due to fragmentation within the sector, because while the opportunities for big data and AI are largely recognised by major tourism stakeholders, smaller companies have not yet identified an action plan to properly adapt to new technologies.

1.2.2.1 Big Data and Sustainable Tourism

Sustainable development has been the most prevalent area of tourism research since the 1990s, and as such, sustainability has caused the most notable impact in the evolution of tourism studies (Ruhanen et al., 2019). As a matter of fact, in a meta-analysis conducted by Ruhanen et al. (2015), a significant increase of 51% was observed in sustainable tourism research publications over the period 2011-2012, showing its recent exponential growth. The reviewed studies centred mainly on the field of sustainable tourism as a general framework, that is to say, "conceptually driven papers focused on processes, practices, theories or trends" (p. 524), usually centred on case studies supported by qualitative and quantitative methods, with the most common approaches being "planning (19%), behavioural studies (12%), perception studies (11%), tourism research theories/methods (10%), indicators and measurement tools (9%), policy studies (9%), and stakeholders (9%)" (p. 526).

Given the importance of sustainability as an area of tourism research, several studies have tried to address the potential applications of big data technologies in sustainable tourism. For instance, Pérez Guilarte & Barreiro Quintáns (2019) examined the main approaches and methodologies for using big data to assess tourism sustainability. In terms of methodologies, the authors consider that the ideal way to strengthen the relation between big data and sustainability would be to incorporate the various data sources (*users, devices,* and *operations)* into open access tourism intelligence systems. In addition, these systems should be connected to conventional tourism sources (surveys, interviews, etc.), and to environmental monitoring systems. However, as their findings suggest, the "measurement of sustainable tourism is perhaps one of the most under researched subjects in tourism statistics, because it lacks the practical tools to guide the implementation and systematicity" (p. 16).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Nevertheless, big data, as applied by Batista e Silva et al. (2018) and Cortina García et al. (2016), can fill this gap by developing specific indicators, especially those contributing to geographical and temporal granularity. These cases show that the integration of big data and indicators to assess tourism sustainability is actually possible, and it also highlights the potential for destination management organisations (Pérez Guilarte & Barreiro Quintáns, 2019).

Similarly, Xu et al. (2019) attempted to define potentially novel methodological areas of application to sustainable tourism studies focusing on the use of big data, namely MNO, GPS, social media, and search engine data. First, large-scale datasets allow for a macro-level study of the natural, economic, and socio-cultural sustainability of tourism, e.g. by analysing cross-country, trans-national or global spatial trends of visitors and their effect on CO2 emissions. Second, various data sources allow different aspects of sustainability to be examined at a local or regional scale: GPS, MNO, or search engine data at a destination can be used to track tourist traffic and to manage over-tourism issues; social media data allows us to explore human relationships with particular locations through sentiment analysis; and administrative data can be used to track biodiversity loss and to control environmental impacts at the destination.

In addition, personalized recommendations and promotions are some of the most powerful attributes of big data. As such, big data technologies could be used to affect the environmental actions of tourists, for example, by leveraging comprehensive analytics to encourage sustainable tourism practices. In terms of tourism sustainability dimensions, while big data has been used to provide insights into tourist movement and carbon emissions, further dimensions are currently being explored with the use of big data, such as issues relating to inequality, food and water security, health and wellbeing, socio-cultural change, clean energy, biodiversity, resource depletion, and climate change (Xu et al., 2019).

Finally, Del Vecchio et al. (2018) attempted to explain how data produced by tourists on social media (namely Twitter, Facebook, and Instagram) and relevant to their travel experiences would support the creation of sustainable tourism experiences. In this case, information generated by tourists on local experiences through the web and social media was useful in identifying trends in tourism satisfaction, critical issues and problems, areas of immediate intervention, opportunities for development and personalisation of tourism products according to gender criteria, geographical location, technological devices, etc. The study reflects on how social big data can promote the

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

process of value development and the competitiveness of destinations through transparent and collective innovation strategies. As a result, social networks can be embraced as tourist networking sites and interactive spaces for exchanging feedback and information, and as repositories of strategic data for companies, institutions, and even tourists.

1.2.2.2 Big Data and Official Tourism Statistics

As previously mentioned, numerous pilot programmes are currently underway, and various statistical agencies have started publishing experimental statistics based on innovative big data sources. The ultimate objective, however, is to transform the tourism statistics system to incorporate multiple sources that are able to satisfy several information needs simultaneously. In a sense, big data is expected to partially replace conventional data sources in the field of tourism statistics. Two different possible cases are illustrated by Demunter (2017) as follows:

a) **Tourism Flows**: MNO data is a clear source for quantifying tourism flows. Call detail records (CDRs) for instance, can provide geographical and temporal details that were previously not available. However, estimations should be adjusted using auxiliary data to counter built-in biases, for example: flight reservation data, to better represent distant destinations where cell phones can be underused when traveling; credit card data; traffic counts; and smart meters. Relevant variables such as trip intention and spending also need to be inferred from other sources, particularly sample surveys. Still, emerging technologies can also contribute to improve data collection, such as integrating automatically collected MNO data with follow-up questions to gather the remaining interesting variables.

b) **Spending**: Credit card data is an appropriate source of information in the case of spending estimations. Point-of-sale (POS) purchases may contain information about the goods or services purchased by the tourists. While data on ATM withdrawals may help estimate cash payments at the destination (although cash brought into the destination country can induce bias, particularly for shorter trips where no local cash withdrawal may be required). Differentiating from non-tourist-related transactions with international companies is essential; overall, it seems highly possible to distinguish between e-commerce and POS transactions. Breakdowns by category of expenditure could be collected from retail cashier data or even traditional surveys.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

As a consequence, improved data analytics will not only allow the use of big data for web marketing, design, and recommendations for tourism, but also to forecast tourist demand and conduct disaster and emergency studies. However, the challenge remains to shift the attention to a "smart" use of big data, by incorporating different layers of knowledge, encouraging real-time use, and ensuring that current trends are properly disseminated. The careful integration of the digital footprints of tourists with industry data would result in a modernization of the tourism data environment. Still, a thorough and integrated tourism data system (comprising statistics, indicators, and big data) remains a priority for the implementation of big data in official tourism statistics (Volo, 2020).

In other words, big data sources of knowledge are truly promising because they can offer reliable and timely information on the tourism sector, producing a more comprehensive spatial analysis than official statistics currently provide (Cortina García et al., 2016). Nevertheless, statistical agencies and data providers must undertake significant and coordinated attempts to obtain results at the same level of quality of existing official statistics.

1.2.2.3 Examples, Applications, and Tools

Substantial efforts have been made in academic research to examine the feasibility of applying big data technologies in the tourism sector. For instance, with regard to data obtained from social networks, Miah et al. (2016) established and analysed a methodology of big data analytics to support strategic decision-making at the destination level. The authors used geotagged images posted by tourists to the social media photo-sharing platform, Flickr, to analyse and forecast tourist behavioural trends in Melbourne, Australia. Four statistical techniques such as text analysis, spatial data clustering, visual information processing, and time series modelling are applied to provide insights into the behaviour of tourists, and to better predict tourist demand. In this case, stable and functional results in both numerical and spatial forms were obtained. As geotagged images are globally accessible, it is possible to evaluate the preferences and behaviour of tourists for inbound travel across multiple markets, as well as clustering lesser-known attractions or less-visited areas of a destination.

Vassakis et al. (2019) explored several methods for the extraction, association, interpretation, and visualisation of Location-Based Social Networks (LBSN) data. In this case, data from Instagram, Flickr, Foursquare and Twitter was applied to the island of

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Crete in Greece, focusing on visitor posts and ratings, nationality, photos, rankings, and engagement. The findings offer particular insights on destination patterns, tourists' sentiment analysis (the highest for typical food and popular tourist locations, and the lowest for the quality of services), and the engagement of consumers across destinations.

In the same way, there is a considerable amount of literature focused on the use of online booking platforms data. For example, Aureli et al. (2013) analysed the quality of online reviews for various three-star and four-star hotels in the Rimini, Italy. In particular, the study discusses whether and to what degree the positive and negative feedback from TripAdvisor impacts online bookings. The authors implemented a content analysis and a panel data analysis. On one hand, the results from the content analysis showed that conventional core services are critical factors in assessing consumer appreciation and criticism (such as room and staff interaction, for instance). On the other, the panel data study seemed to show a linear association between hotel success and online credibility.

Xiang et al. (2015) tried to explore and illustrate the use of big data analytics in relation to hospitality issues, including the relationship between hotel guest experience and satisfaction. Specifically, the researchers applied a text-analytical method to a vast number of user reviews extracted from Expedia.com to deconstruct hotel guest interactions and examine the correlation with satisfaction scores. This research showed the usefulness of big data analytics to detect novel trends in hospitality using consumer-generated information that is readily accessible online. Although the results were based on data generated on a single website and over a specific time span, they reflected how users were communicating about their experiences in online hotel reviews.

Similarly, Joseph & Varghese (2019) applied text mining techniques to Airbnb user review data from London, England, with the purpose of determining the drivers of customer satisfaction. In particular, the most common positive aspects within Airbnb reviews were related to the location, room amenities, comfort, cleanliness, staff, food and drinks, quietness, beds, value for money, customer support, design, facilities, view, Wi-Fi, and payment. These findings seem to be consistent with the determinants of customer satisfaction in the hotel industry. For instance, cleanliness, service quality, and employees' knowledge and service are considered the most critical aspects in determining hotel guest's satisfaction.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Batista e Silva et al. (2018) aimed to develop the existing knowledge base of the current spatiotemporal distribution of tourism in the EU-28 in order to provide new perspectives and applications related to tourism management and policy. Specifically, the authors sought to improve the geographical specificity of current data on the spatial distribution of tourism demand at regional level; extract regional temporal profiles of tourism demand; generate high-density tourist maps on a monthly basis; and leverage the generated data to determine regional tourism dimensions such as tourism intensity, seasonality and vulnerability. For this purpose, data was integrated from a variety of sources, namely European official statistical organisations, and online booking services. As a result, a novel, complete, and reliable dataset was generated detailing the average daily number of overnight tourists in different EU regions, and their spatial resolutions with monthly breakdowns. The constructed dataset made it possible to discern the main spatial-temporal trends and characteristics of tourism in Europe on a regional and local scale.

Notably, Xiang et al. (2017) provided a comparative evaluation of three main online review sites, namely TripAdvisor, Expedia, and Yelp, in terms of information quality for the entire hotel population in Manhattan, New York, USA. The results indicate that there are major gaps in the representation of the hotel industry on these platforms. Online reviews especially differ considerably in their language characteristics, semantic characteristics, sentiments, ranking, usefulness, and even the relationship between these features. In terms of the sheer volume of review results, TripAdvisor and Expedia are similar to each other, while Yelp contains a significantly lower number of reviews. Furthermore, between TripAdvisor and Expedia, which are similar in the amount of reviews, the former tends to have a better overall quality. However, TripAdvisor and Yelp tend to present reviews which could be perceived as more useful than those presented by Expedia. These results indicate that one platform alone may not be an appropriate source of quality data since multiple platforms could have relatively unique features.

Other valuable platforms have also been examined in relation to tourism research on big data. Particularly, Ferreira Dinis et al. (2019) provided a framework for constructing composite indicators to assess the interest of tourism destinations using online searches. This methodology was applied to quantify the online search interest for Portugal by international tourists, in particular from countries such as Spain, the United Kingdom and Germany. Data from Google Trends were used to create composite indicators based on the statistical structure of the Tourism Satellite Account (TSA) for the following categories: Food and Beverage Services, Accommodation Services,

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Passenger Transport Services and Transport Equipment Rental Services, Travel Agencies and other Reservation Services, Cultural Services, Sports and Recreational Services, and Miscellaneous Tourism Services. The composite indicators displayed correlations with the effective tourist demand of these markets in Portugal, which suggests that indicators derived from Google Trends data can act as a proxy for inbound tourism in Portugal.

Equally important, Valcke (2019) demonstrated the use of big data for crisis monitoring in tourism in Flanders, Belgium. In particular, the research focused on how Visit Flanders used website monitoring and flight data for different market segments, as a result of the 2015 and 2016 terrorist threats. Since Visit Flanders normally uses the Synthesio web monitoring system, it was relatively simple to collect data from online social media and booking sites. Specifically, the central web monitoring approach was based on the amount of discussion regarding insecurity in relation to Flanders. The results managed to illustrate how different international tourism markets were responding to Flanders during the crisis period, using social media data and flight reservation.

Furthermore, Fuchs et al. (2014) outlined the knowledge infrastructure recently established in Åre, one of the most emblematic ski areas and resorts in Sweden. The Destination Management Information System Åre (DMIS-Åre) uses a Business Intelligence approach to guide the development and implementation of knowledge as a precondition for operational learning at tourism destinations. The pre-trip and post-trip phases are the focus of the system, in this sense, customer-based knowledge channels such as tourist search (web navigation), booking, and feedback behaviour (surveys, online review platforms) are the foundation of the DMIS-Åre, a platform that involves all the relevant tourism stakeholders in the area.

Finally, several studies have been carried out on the use of MNO positioning data for tourism statistics since 2008. In their seminal paper, Ahas et al. (2008) introduced the potential of passive mobile positioning data for the study of tourism. They used a database of roaming locations and call activities in network cells. Using examples from Estonia, they identified the particularities of MNO data, the process of collecting the data, sampling, accurately handling the spatial database, and appropriate analysis methods. The findings showed that MNO data has significant applications in for geographical studies, since in this particular case the correlations with traditional accommodation statistics in Estonia was up to 0.99 in the most regularly visited tourist

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

areas. Similarly, Yamamoto (2019) attempted to identify the number of tourists in various Japanese destinations, along with their demographic characteristics, by analysing the location data of mobile phone users collected by MNOs. As an illustration, Table 3 provides a general overview of the most commonly used tools in the previously mentioned examples of big data for tourism.

Table 3. Commonly used tools in tourism big data initiatives

Name	Description	Licence
R	Programming language with a focus on statistical analysis	Free software / GNU General Public License, v2
Python	General purpose programming language	Free software / Python Software Foundation License (PSFL)
Anaconda	Distribution of Python and R for data science	Free software / New BSD License
Apache Hadoop	Framework for processing and storage of big data applications	Free software / Apache License 2.0
Apache Spark	Data processing framework for big data applications	Free software / Apache License 2.0
Apache Hive	Data warehouse software for data query and analysis	Free software / Apache License 2.0
TensorFlow	Software library for machine learning	Free software / Apache License 2.0
Weka	Machine learning software	Free software / GNU General Public License
Elasticsearch	Search and analytics engine	Free software / Apache License 2.0
Kibana	Data visualization dashboard for Elasticsearch	Free software / Apache License 2.0
Pentaho	Business intelligence software	Free software / Apache License 2.0
Tableau	Data analysis and visualisation software	Proprietary
Power BI	Business intelligence software	Proprietary

Overall, the implementation of big data in the tourism sector is still at an early stage. However, its clear potential to supplement tourism statistics has been widely recognised, as demonstrated by the current initiatives to develop methodologies for the integration of big data sources into official statistics. As previously stated, the ideal solution is to transform the current state of tourism statistics into a system that incorporates a wide range of both traditional tourism survey data, as well as big data sources.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

1.3 ESSnet Big Data II

Within the European Statistical System (ESS), several initiatives have been developed concerning the application of big data in official tourism statistics. Namely, the adoption of the Scheveningen Memorandum, which calls for an action plan concerning big data and official statistics, and the subsequent creation of task forces on big data within the European Statistical Office (Eurostat) and the European Statistical System (ESS). Further initiatives consist mainly of pilot projects to develop in-house technological expertise. Nonetheless, the most ambitious initiative so far is undoubtedly the European Statistical System's ESSnet Big Data project. ESSnet is a project launched by Eurostat in 2016 to study the potential of big data sources to produce official statistics. These pilot projects include web scraping (for company characteristics and job vacancies), smart meters, mobile phone data and AIS data (automated tracking of ships) (Demunter, 2017).

The ESSnet Big Data I project was conducted from February 2016 to May 2018 with the participation of 22 EU partners as an initial effort to prepare the ESS for big data integration. Its main goal was to incorporate big data into the daily output of official statistics, by exploring the potential of selected big data sources through pilot projects and concrete applications (European Commission, 2020).

Specifically, the project was divided into 10 work packages (WP): WP0 Coordination, WP1 Web scraping job vacancies, WP2 Web scraping enterprise characteristics, WP3 Smart meters, WP4 AIS Data, WP5 Mobile phone data, WP6 Early estimates, WP7 Multiple domains, WP8 Methodology, and WP9 Dissemination.

Subsequently, the ESSnet Big Data II project started in November 2018 as a continuation of ESSnet Big Data I. The ultimate purpose of the ESSnet Big Data II project is to better prepare the ESS for the incorporation of big data sources into the development of official statistics, considering the results from the previous project. The project incorporates the following specific objectives and actions:

a) **Methodology:** The main focus of the initiative lies in the development of new pilot projects that apply the previously developed methodology, with a particular emphasis on the different approaches for combining data sources. The goal is to deliver results in a more user-oriented manner than in the previous ESSnet by offering solutions to potential users who are dealing with specific problems related to big data sources.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

b) **Quality:** In this case, the objective is to establish a specific quality structure for the use of big data sources based on the quality system already established by the ESS.

c) **Process and Architecture:** When applying the findings of the previous ESSnet, statistical processes need to be developed within an architectural context. Components will be defined depending on the source of big data and the domain of use, and consideration will be taken for broader use in the ESS to improve the relationship with international statistical standards. In particular, the European Commission's *Big Data Test Infrastructure (BDTI)* initiative is particularly relevant to the project, since it provides EU public administrations with a free cloud-based analytics test environment to experiment with big data technologies, as well as the possibility to prototype big data solutions before implementing them on their premises. A variety of open source solutions, data sources and the necessary cloud infrastructure comprising virtual machines, analytics clusters, storage facilities and networking facilities are part of the test environment provided by BDTI (European Commission, 2019).

As summarized in Table 4, the ESSnet Big Data II project is comprised by 11 work packages: WPB Online job vacancies, WPC Enterprise characteristics, WPD Smart energy, WPE Tracking ships, WPF Process and architecture, WPG Financial transactions data, WPH Earth observation, WPI Mobile networks data, WPJ Innovative tourism statistics, WPK Methodology and quality, and WPL Preparing smart statistics.

In general, the ESSnet Big Data II project intends to further prepare the ESS to successfully incorporate big data sources into official statistics. Given the characteristics and diversity of big data sources, the project entails not only their potential uses in official EU statistics, but also the resolution of any possible challenges and problems related to their implementation.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Table 4. ESSnet Big Data II Work Packages

Work Package	Description
WPB Online job vacancies	The goal of this WPB is to generate statistical estimates for online job vacancies. Specifically, WPB focuses on the use of web scraping techniques to generate accurate data from career portals, job advertisements on business websites, and data from third-party sources on job vacancies.
WPC Enterprise characteristics	WPC deals with the use of web scraping, text mining and inference techniques to collect and process enterprise information with the purpose of enhancing or updating existing information in the national business registers, such as Internet presence, operation type, address information, ownership structure, etc.
WPD Smart energy	The purpose of WPD is to incorporate smart meter data in the production of statistical indicators. In this sense, it involves linking electricity data to other administrative sources, to produce statistics related to enterprises, households, and accommodation services.
WPE Tracking ships	WPE aims at developing functional production prototypes, including the establishment of procedures and the development of technological solutions, to promote and support the collection, processing, and analysis of big data from AIS (Automatic Identification System).
WPF Process and architecture	WPF seeks to define the reference architectures required for the development of big data at both national and European levels, with the purpose of providing concrete solutions on which National Statistics Institutes (NSIs) can rely for their own big data production.
WPG Financial transactions data	The key objective of WPG is to gain an overview of the potential sources of financial transaction data and the data infrastructure. The goal is to explain the accessibility and statistical potential of the available financial transaction data.
WPH Earth observation	Official statistical production and landscape mapping are some of the most noteworthy applications of Earth Observation (EO) data. As a result, the purpose of WPH is to integrate EO data from different sources that can be useful for a wide range of thematic areas, such as agriculture, construction, land and settlement cover, and forestry.
WPI Mobile networks data	WPI focuses on the improvement of a production framework already introduced in previous ESSnet projects. Still, the overall objective is to develop standardised statistical production processes for the collection and use of mobile phone data.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

WPJ Innovative tourism statistics	By developing a pilot project for a Tourism Information System, WPJ intends to propose a conceptual framework for the use of big data in tourism, which will promote statistical development in the field of tourism by combining different big data sources with administrative registers and statistical databases.
WPK Methodology and quality	The purpose of WPK is to consolidate all the relevant information acquired during the implementation of both ESSnet Big Data projects regarding the appropriate methodologies and quality guidelines for the use of big data in official statistics.
WPL Preparing smart statistics	The goal of WPL is to explore the statistical potential of data related to the widespread use of smart wear, city sensors, vehicle sensors, and other smart systems. Specifically, WPL deals with the production of smart statistics in the context of agriculture, cities, devices, and traffic.

Note. Adapted from "ESSnet Big Data" by the European Commission, 2020a (https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data). CC BY-SA.

1.3.1 ESSnet Methodology

One of the most important aspects of the ESSnet Big Data II project is the consolidation of appropriate methodologies for the use of big data and the development of quality guidelines. In the context of the project, the suggested methodologies include only those who fulfil the following characteristics:

a) Methods used by researchers involved in the ESSnet Big Data I and II projects.

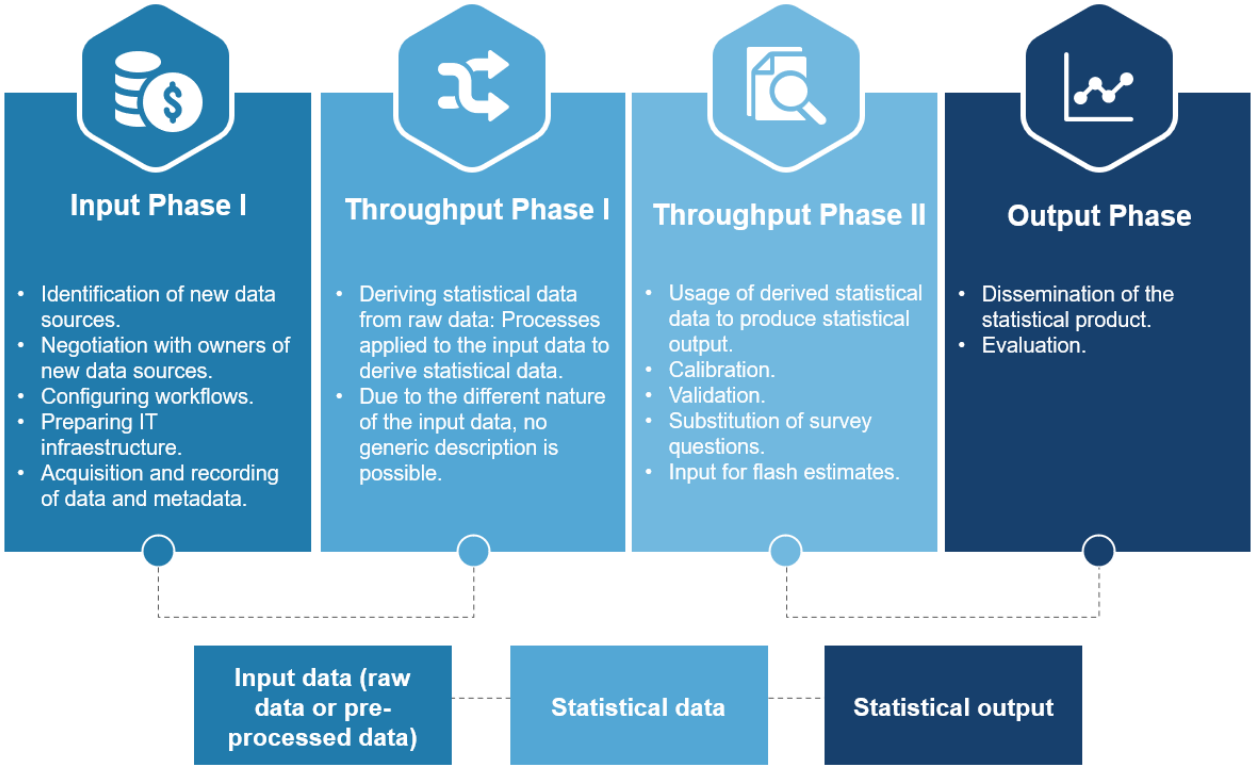
b) Methods included in one of the big data official statistics currently published, or in the process of becoming officially published (Eurostat, 2020e).

In general, the methodology follows a production process logic, with a particular focus on the specific phases that are affected by the use of big data sources (input phase, throughput phase I, and throughput phase II), as detailed in Figure 5.

Regarding the **input phase**, the potential uses of big data in official statistics have significantly encouraged the implementation of new data sources, in particular, text and image data are increasingly being used. As a consequence, text mining and NLP methods such as text categorization, text clustering, and sentiment analysis; as well as image mining methods, including bag of visual words and deep learning, have become an essential part of the methodology used for the production of official statistics.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Figure 5. Phases of the production process with big data sources



Note. Adapted from "Deliverable K3: Revised Version of the Quality Guidelines for the Acquisition and Usage of Big Data" by Eurostat, 2020, *WPK: Methodology and Quality*, p. 6.

Furthermore, with respect to the **throughput phase I**, the most relevant processes are related to the combination of multiple data sources (linking methods) and dealing with errors arising from the use of big data.

After the collection of data during the previous phases, **throughput phase II** focuses on the production of statistical output. As a result, the project outlines several uses of big data sources in official statistics: Big data sources as input for the production of official statistics, replacement of questions from surveys, validation and comparison of results with results from traditional data sources, and flash estimates based on leading or correlated indicators (Eurostat, 2020d). In particular, two different cases of the use of big data sources in official EU statistics have been identified during the ESSnet Big Data II project: the calculation of the Consumer Price Index that incorporates prices of products scraped from the web and scanner data (currently implemented in Austria, Belgium, Denmark, France, Italy, Luxembourg, the Netherlands, Norway, Sweden and Switzerland), and the Traffic Intensity Statistics of Statistics Netherlands that uses road

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

sensors to produce vehicle counts data. The rest of the initiatives identified are still of an experimental nature.

1.3.1.1 Linking Methods

Multiple alternative sources are increasingly used together by statistical offices for the development of statistical indicators. This method not only makes it possible to increase the otherwise available information from a single source, but it also improves the accuracy of each single source, thus increasing both the quality of its measurements and its coverage in terms of statistical target units. In addition, when a single source does not provide complete data, a combination of sources is a common technique for addressing the problem. As a matter of fact, by measuring the same variable across multiple data sources, it is possible to identify and correct errors and to harmonize variable definitions. In this sense, micro-integration is the most commonly used approach by statistical offices, followed by alternatively modelling the measurement errors (Eurostat, 2020e).

The combination of sources is possible at unit or aggregate level. In this case, when specific unit identifiers are available, such as unique personal identification numbers, deterministic linkage may be used; otherwise probabilistic linkage could be used. However, when a combination of sources is carried out through linking units, such as adjusting for single source coverage or selectivity, the problem is especially complex when dealing with big data, since big data records are generally different from the target statistical units and contain very little information that could distinguish them (Eurostat, 2020e).

As explained by Eurostat (2020e), “the methodology to execute deduplication and linking activities is mainly the record linkage. The method used is often based on deterministic decision and sometimes on machine learning classification techniques (logistic regression, classification trees)” (p. 27). In this case, the most common recommendation arising from the project is the need to include record linkage and business registers specialists at National Statistical Institutions (NSIs), considering that data linking knowledge combined the understanding of company structures play a crucial role in generating statistical outputs.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

1.3.1.2 Errors

A classification of potential errors with generally agreed descriptions already exists for survey-based statistics. In the case of multi-source statistics, it is far more challenging to provide an exhaustive list of potential errors with their corresponding category, as a result of the diversity among big data sources However, the ESSnet Big Data II project has identified a list of the most relevant errors when handling big data sources, as explained by Eurostat (2020d):

a) **Coverage**: The degree of control NSIs have on data collection is among the main differences between big data and conventional sources. Because data generation depends on external factors and not on the decisions of the NSIs, it is likely that such data may not be reflective of the entire population, but only a fraction of it, resulting in possible sampling bias. For instance, when taking Twitter data into account, it is clear that the tweets collected apply only to particular subsets of the more general population: the subset of individuals with a Twitter account and the subset of Twitter users who have chosen to post publicly.

b) **Comparability over time**: The data generation process is one of the key issues about the integration of big data into official statistics. Big data is generated and supported by independent actors for a non-statistical purpose. Besides the possible introduction of coverage errors at a fixed time point, an unregulated process can cause problems with the comparability of data between two or more reference periods over a longer time span. These issues primarily depend on the data structure's consistency over time.

c) **Measurement errors**: The difference between the true value and the value obtained during the measurement process is called measurement error. In the case of survey-based statistics, measurement errors are commonly attributed to human factors. But when it comes to big data sources, measurement errors are not as common since they are mostly related to faults in the instruments for data recording.

d) **Model errors**: Statistical models have the special property of having a foundation in probability theory, while machine learning algorithms are usually of an empirical nature. For classical statistical models, like linear regression, and also

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

for advanced machine learning algorithms, such as random forest or deep learning, model errors can occur simply by not specifying an important variable.

e) **Processing errors**: The processes involved in generating statistical outputs using big data sources are as diverse as the big data sources themselves, therefore, the potential processing errors can be quite varied. However, processing errors usually involve data entry errors, coding or mapping misclassifications, editing and imputation errors, unification errors, unit errors, and linkage errors, but in the context of big data manual activities play a minor role.

Besides a preliminary classification of errors related to the use of big data sources, the ESSnet Big Data II project provides the following recommendations to manage these potential challenges:

> Clearly establish the population of interest.
>
> Short surveys may be launched in order to identify the characteristics of an observed population.
>
> Closely monitor the structure of the data. Check each data generation on structural changes in comparison to the previous one.
>
> Rely the statistical output on more than one source of data. The sources can be of different typology: big data, administrative data, survey data.
>
> Apply appropriate model selection and evaluation criteria. Techniques like cross validation, out-of-sample tests, etc. should be applied wherever possible to assess the model quality and possible errors.
>
> Compare multiple machine learning and statistical models. Since it is not always straightforward to choose the right tool for the job, different methods should be tested and evaluated.
>
> Evaluate the bias of the training data set. In supervised learning, an unbiased training data is very important to not estimate based on a biased model. (Eurostat, 2020d, pp. 26–28)

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

1.3.2 ESSnet Quality Guidelines for the Acquisition and Usage of Big Data

Big data sources do not only vary from conventional data sources in terms of structure and content, but also in terms of how data is acquired or collected. In this case, there are two possibilities: direct cooperation with the data provider, or the collection of publicly available data on the web through web scraping techniques or APIs.

Under these conditions, Eurostat (2020d) has established a set of principles for any procedure of accreditation of non-official data sources:

> Principle 1: Accreditation must be fully compliant with the well-established principles and quality frameworks that guide the world of official statistics and consistent with quality assurance practices embedded deeply in the work of statistical offices.
>
> Principle 2: Any accreditation procedure must be flexible in a way that does not unduly prejudice or rule out new opportunities without serious examination.
>
> Principle 3: An accreditation procedure should include sequential decision-making based on a pragmatic stepwise approach, so that new data sources that will not work are spotted early on, while investment in those that will work is not jeopardized.
>
> Principle 4: The accreditation procedure must contain an empirical assessment with real data and it must be carried out by statistical offices directly. It cannot be delegated to filling out questionnaires by the source owners.
>
> Principle 5: A systematic accreditation procedure must assess the quality of the statistical inputs (including the source and metadata), of the statistical outputs, as well as of the statistical processes involved.
>
> Principle 6: The final decision for the accreditation of a new data source must also incorporate a combination of corporate criteria, broader than strict data quality. The accreditation procedure must compile adequate supporting documentation, including measurements. (pp. 11-12)

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

1.3.2.1 Data Sources – Cooperation

In the case of direct cooperation with the data provider, big data sources pose new concerns regarding governance and responsibilities. In order to efficiently process big data from external sources, it is necessary to develop a consistent governance framework and a clear distribution of responsibilities. Therefore, within the NSI, an individual or organizational unit should be assigned to be responsible for searching and acquiring new sources. This approach has already been implemented by Statistics Netherlands, under the denomination of "Data Scouting", a specific job role that encompasses functions such as improving relationships between internal and external stakeholders; defining data needs with internal and external users; exploring possible relevant new data sources; and negotiating conditions, business models, and agreements for the use of big data sources (Eurostat, 2020d).

As a result, the following guidelines have been proposed by the ESSnet Big Data II project:

a) As the NSI gains knowledge of a new, potentially useful big data source, it is important to notify all departments within the NSI that may have an interest in using the new data.

b) The new data sources delegate should collect the exact necessary information, by examining the following areas: the population coverage, the units of measurement, variables, timeliness and frequency, and information on the organisation.

c) Data file requirements must be addressed in a professional manner, in particular the means for data access, and in the case of pre-processed data access, the clarity of the technical processes applied to the data, the transmission time, and the metadata should be clearly specified.

d) During the forensic investigation of the test data, it should be clarified which technical and statistical processes are required for using the new data source, if the skills necessary for data processing are available in the statistical office, and if the new data can be adequately managed by the statistical office's available resources.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

e) Long-term access to the big data source must be guaranteed in order to avoid coverage and compatibility errors.

f) Governance issues, including a change in management and a dispute resolution mechanism, need to be addressed (Eurostat, 2020d).

In this context, some of the most promising sources of big data are Automatic Identification System (AIS) data, smart meter data, and Mobile Network Operator (MNO) data. Particularly, the use of MNO data has been extensively reviewed as part of the ESSnet Big Data II project, since MNO data offers information about the geolocation of individuals, internet traffic, and interactions between individuals. Consequently, a set of guidelines have also been established in relation to the acquisition of MNO data, summarised as follows:

a) **Agree on roles:** Agree with MNOs on the different roles played by MNOs and the NSI. The NSI should be part of the design of the whole end-to-end statistical process.
b) **Audit raw data extraction:** Agree with MNOs on the raw data to be used in the statistical process. This should be the result of a trade-off between the adaptation of the ESS Reference Methodological Framework for the Production of Official Statistics with Mobile Network Data and the technological and business feasibility to use this data.
c) **Audit raw data pre-processing:** Agree with MNOs on the statistical processing of raw data to generate intermediate data for the further statistical analyses.
d) **Document data provenance and pre-processing:** Document both the data provenance (which raw data exactly to use) and the data pre-processing (generation of intermediate data: method, parameters, etc.). This documentation must find an optimal trade-off between public transparency, privacy and confidentiality, and industrial secrecy and intellectual property rights. (Eurostat, 2020d)

1.3.2.2 Data Sources – No Cooperation

When direct cooperation with the data provider is not feasible, there are several methods for collecting or recording big data. In general, for official statistics purposes, web scraping is the easiest and least costly way to obtain big data. However, it is necessary to ensure that the raw data is readable and linkable to other formats in order

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

to obtain high quality data. With this in mind, the ESSnet Big Data II project has established the following web scraping guidelines:

> Ensure that each data set will have a corresponding metadata set. Use the unified format for data and metadata store.
>
> When collecting the data, ensure that there are reliable attributes that can be used to link to other data (e.g., geolocation, NACE, etc.).
>
> If possible, allow access to the raw data with the unified interface, i.e. the same name of fields for the specific dimension, e.g. company_id, NACE.
>
> If there are any methodological differences in the interpretation of the same dimension, e.g. job vacancy vs. job offer, please use the metadata.
>
> Ensure that all data is stored in a secure way and try to create different groups of users, e.g. external users vs. internal users to allow limited access to the data.
>
> Try to estimate the target population size, if possible, and use metadata to store this information.
>
> For web scraping, follow the document "ESS web-scraping policy" prepared by ESSnet Big Data WPC.
>
> Use similar classifications, if possible, or at least create the transition key to encode/decode the list of possible values from one data source to another.
>
> Store the data in machine readable format, which can be processed by the computer.
>
> If possible, allow access to raw data in standard formats like JSON or CSV, to be easily loaded into most common data science environments, like Apache Hadoop, Python or R. (Eurostat, 2020d, pp. 15–16)

Similarly, in the case of social media data, a comprehensive set of guidelines have been developed. First, the population of interest needs to be established and carefully researched. The social media platform should be selected on the basis of stability, both in terms of time (data should be available over an extended period of time), access

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

policy, and algorithm changes. And finally, it is crucial to continuously review the implemented techniques to capture any changes in the structure of the data (Eurostat, 2020d).

Overall, an extensive study of the use of big data sources has been carried out by the ESSnet Big Data II project, and several promising data sources have been identified. Preliminary results suggest that the incorporation of big data into official statistics is still in a state of development, with particular issues pertaining to data access, availability, protection, confidentiality, linkage, and processing. Altogether, the use of big data offers many possibilities to enrich official statistics, and the project has already tried to solve certain methodological problems through the creation of guidelines and methodological recommendations.

1.3.3 WPJ - Innovative Tourism Statistics

One of the focal points of the ESSnet Big Data II project is centred on innovative tourism statistics. Specifically, the Work Package J (WPJ) aims to resolve the need for a conceptual framework in tourism big data by creating a smart Tourism Information System. In this way, the system seeks to incorporate innovative statistical techniques to integrate multiple big data, administrative, and statistical sources related to tourism. The WPJ is currently being undertaken by the following institutions: GUS (Statistics Poland), NSI (Statistics Bulgaria), ELSTAT (Statistics Greece), HSL (Hesse Statistical Office, Germany), ISTAT (Statistics Italy), CBS (Statistics Netherlands), INE (Statistics Portugal) and NSI (Statistics Slovakia). So far, the WPJ has been able to present preliminary results concerning the potential inventory of data sources and web scraping procedures to be implemented, as well as an overview of the methods related to the combination and spatiotemporal disaggregation of data from different sources (Eurostat, 2020a).

Specifically, the WPJ has established the following series of tasks to be conducted during the duration of the project, as explained by Eurostat (2020a):

a) Inventory of big data sources related to tourism statistics: Containing information collected with web scraping techniques regarding accommodation establishments, transport, traffic flow in local communication, admission tickets, meteorological data, communications on epidemiological threats and natural disasters, etc. As well as external data such as water consumption, waste

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

production, energy meters, parking, traffic, store transactions, payment cards, mobile data, and so forth.

b) Examining the availability, legal aspects and the quality of the new identified data sources used in the project.

c) Developing a methodology for combining and disaggregating data from various sources.

d) Flash estimates in the field of tourism: Flash estimates refer to readily available data without the usual long delay observed in official tourism statistics. In this sense, the project seeks to integrate big data sources with sample surveys to determine future movements of tourists and their expenses.

e) Use of big data sources and the implementation of the developed methodology to improve the quality of data in various statistical areas: The estimated data will be employed to verify the size of tourist traffic according to travel directions, means of transport, types of accommodation, and tourist expenditure aggregates. Additionally, the Tourism Information System will serve as a basis to improve the existing Tourism Satellite Accounts.

f) Description of challenges and recommendations on legal aspects, availability and sustainability, methodology, quality, and technical requirements to support the implementation of big data in tourism.

Additionally, during the first stages of the project, a series of current initiatives concerning the use of big data in official tourism statistics were identified. For the most part, NSIs are focusing their efforts on establishing direct cooperation with MNOs to obtain anonymous data from mobile networks (Hesse Statistical Office, Germany; Hellenic Statistical Authority, Greece; Statistical Office of the Slovak Republic, Slovakia; National Institute of Statistics, Italy), as well as with financial entities which possess data on financial transactions in the tourism sector (Hesse Statistical Office, Germany), and the use of traffic sensors and traffic images data (Statistical Office in Rzeszów, Poland; National Institute of Statistics, Italy).

However, these efforts are still in the process of being developed. Since as previously mentioned, obtaining big data from external sources through direct cooperation with data providers is a lengthy process, especially in the case of tourism statistics, because it requires “establishing cooperation with the data administrators, determining

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

conditions of access, concluding contracts and conducting arrangements regarding the scope and frequency of data acquisition" (Eurostat, 2020a, p. 9).

In general, WPJ uses as a basis the methodology and quality guidelines developed during the ESSnet Big Data I and II projects and adapts them to the particular case of using big data sources within tourism statistics.

1.3.3.1 Inventory of Big Data Sources Related to Tourism Statistics

As one of the main goals within WPJ is the integration of external and internal data sources, an inventory of the potentially useful data sources was created. The data sources identified were then divided according to their suitability for estimating tourism demand and supply, and as a result it was found that most sources can only be used to assess tourism demand. With regard to external data sources, the findings of the inventory showed that external data sources include those not yet accessible to partner countries as well as those with limited availability, nonetheless, in all partner countries 52% were found to be accessible. But the highest proportions of available external sources were observed in Poland (around 85%) and Portugal (83%).

Still, because a great amount of information useful for the estimation of tourism statistics is publicly available online, WPJ has focused a large part of its efforts on the creation of guidelines for web scraping of online tourism portals. As a result, a catalogue of online tourist portals with potential statistical use was identified based on web pages with the following characteristics: "a large amount of indexed data, high detail of data and frequent updating, high activity and interaction of users, [and] usability to enrich the results of traditional statistics" (Eurostat, 2019a, pp. 7–8). An overview of the selected online portals can be found in Table 5.

Table 5. Classification of selected online tourism portals by available information

Type of Information	Portal
Accommodation (supply and demand)	Hotels.com Booking.com Kayak TripAdvisor Airbnb Zoover

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Transport	Booking.com Kayak TripAdvisor HolidayCheck Idealo
Food and Beverage Services	TripAdvisor Zomato

Note. Adapted from "Deliverable J1: Methods for webscraping data processing and analyses" by Eurostat, 2019, *WPJ: Innovative Tourism Statistics*, p. 9.

These online portals were evaluated in terms of the number of offers, available variables, and data accessibility. In this case, variables such as the type of accommodation and number of beds have been identified as significant to the tourism supply side. While the prices for accommodation, transport, restaurants, and other services have been identified as relevant to the tourism supply side.

First, Hotels.com and Booking.com were selected during the first stages of the project to develop and test web scraping scripts that could be implemented by the partner countries. In the following stages of the project, it is planned to further evaluate other sites containing data on tourist accommodation (Booking.com, Kayak, Airbnb), as well as on visitors' travel and expenses related to transport, food, and other tourist services (Expedia, TripAdvisor, and 365 Tickets) (Eurostat, 2019a).

Since big and administrative data are not generated for official statistical purposes, possible consistency, methodological, legal, and privacy concerns should be clearly identified and resolved before incorporating new sources into the production system of official statistics. Under these assumptions, the WPJ inventory for big data sources contains four different aspects, variable identification (detection of relevant variables), variable taxonomy (classification of variables into topics), variable mapping (relationships between variables from external and internal data sources), and variable ontology (subset of core variables).

1.3.3.2 Flow Models

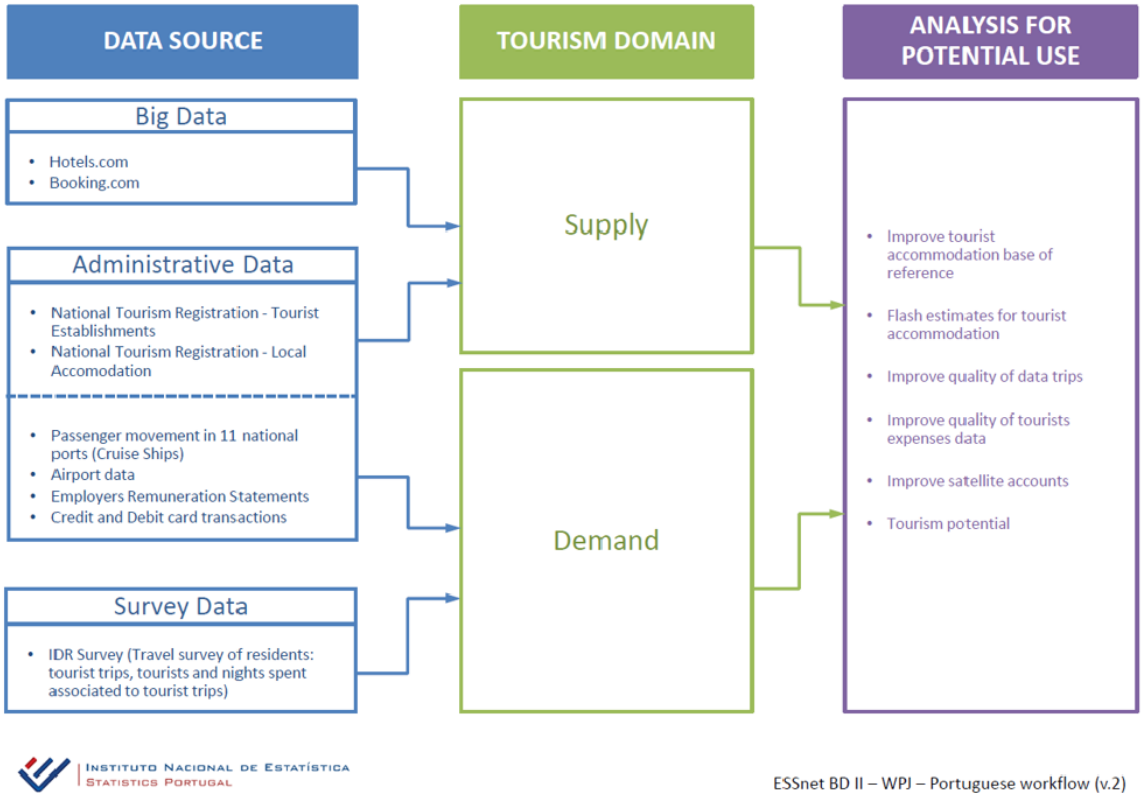
Based on the data sources inventory, the project partners have developed flow models according to the agreed scheme for their countries. Each model provides suggested guidelines for integrating information obtained from external sources and web scraping with official statistical data, in relation to each particular country. The models developed by the project partners have been adapted to the number and types of sources identified

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

in each country and to the areas in which they will be applied in the short term (Eurostat, 2020a).

As an example, the flow model for the particular case of Portugal can be visualized in Figure 6. Specifically, tourism supply estimates will be compiled using data collected through web scraping techniques and data from Portugal's National Tourism Register. As a result of the combination of data from these sources, six anticipated results are planned, including improving the quality of data on tourist spending, improving the quality of travel data, flash estimates for tourist accommodation and improving the Tourism Satellite Account (TSA).

Figure 6. Flow Model for Portugal



Note. Reprinted from "Deliverable J2: Interim technical report showing the preliminary results and a general description of the methods used" by Eurostat, 2020, *WPJ: Innovative Tourism Statistics*, p. 45.

Consequently, through the experience gained during the development of the various flow models for each partner country, it was then possible to consolidate a common set of proposed guidelines and areas of improvement:

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

a) **Improving tourist accommodation data**: The use of web scraping methods and geographical coordinates will improve both the completeness and quality of results for tourist accommodation establishments.

b) **Spatial disaggregation of tourist accommodation data**: Data obtained from non-statistical sources (administrative and big data) will allow to obtain reliable information about tourist accommodation establishments at the lowest levels of territorial division (villages, districts, housing estates, etc.).

c) **Flash estimates of tourist accommodation data**: Flash estimates in the field of tourism will respond to the growing demand from stakeholders regarding the rapidly changing situation on the tourism market.

d) **Improving the quality of data on trips**: The new data sources and statistical methods allow to verify tourist traffic estimations according to travel directions, means of transport, types of accommodation.

e) **Improving the quality of expenditure data**: The use of data on prices of accommodation, restaurants, transport, and ticket prices for tourist attractions allow to verify tourist expenditure estimations.

f) **Improvement of Tourism Satellite Accounts**: By improving the data quality for trips and tourist expenditure, it is then possible to better calculate the direct contribution of tourism towards the overall economy.

g) **Urban tourism**: The information obtained from Smart City systems can be used to estimate changes in the urban population at a given point in time. Furthermore, information on the number of arriving and departing vehicles can help in determining the place of origin of a vehicle (country and abroad) as well as in estimating the number of people visiting the city.

h) **Event related tourism**: Based on new data sources it will be possible to estimate the number of people traveling in relation to fairs, conferences, etc.

i) **Tourism potential**: Detailed, high-quality data on accommodation, infrastructure, and natural resources will allow a more complete picture of the regions' tourism potential. At the same time, better estimates of the number of tourists and their

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

expenditures will make it possible to determine more precisely the utilization of the tourism potential.

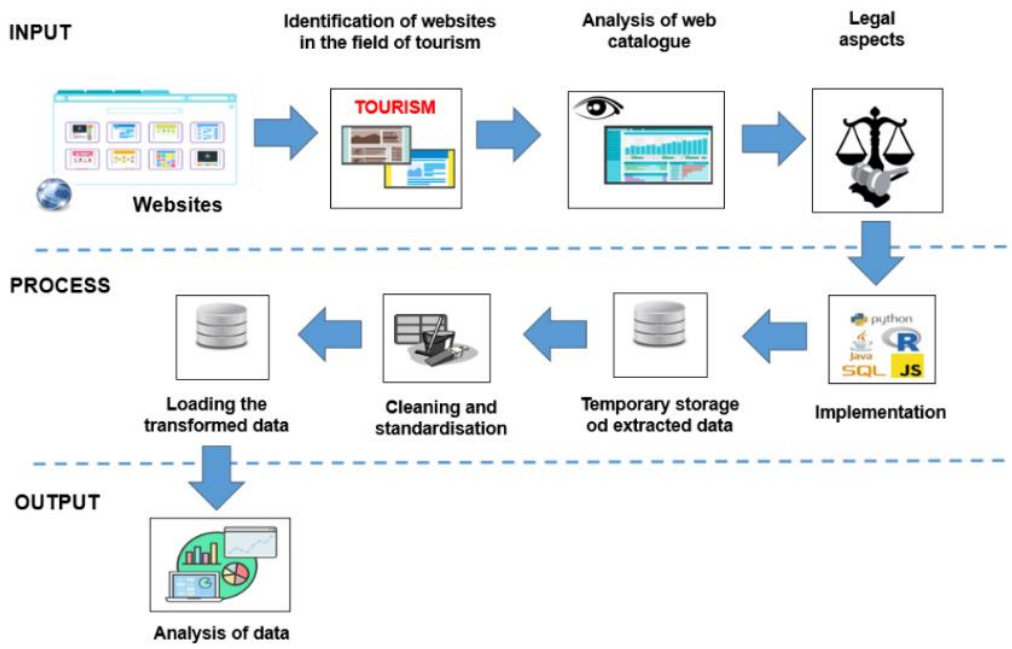
j) **Tourism attractiveness**: Data obtained from various sources at different levels of territorial divisions will provide the necessary information to develop indicators determining the tourist attractiveness of regions and cities (Eurostat, 2020a).

1.3.3.3 WPJ Web Scraping Tool

As part of the WPJ, development began on a universal, automated tool for retrieving data from unstructured online sources related to tourism. The proposed solution will enable a person without prior knowledge of web scraping to retrieve online data. In this way, the tool will allow to easily create a web scraper through a user-friendly interface (Eurostat, 2020a).

In order to gather data from different online platforms, WPJ has established its own notion of a web scrapping methodological process for the field of tourism. It is organized into three different stages, as illustrated in Figure 7.

Figure 7. WPJ web scraping process



Note. Reprinted from "Deliverable J1: Methods for webscraping data processing and analyses" by Eurostat, 2019, *WPJ: Innovative Tourism Statistics*, p. 6.

Project co-financed by the European Regional Development Fund

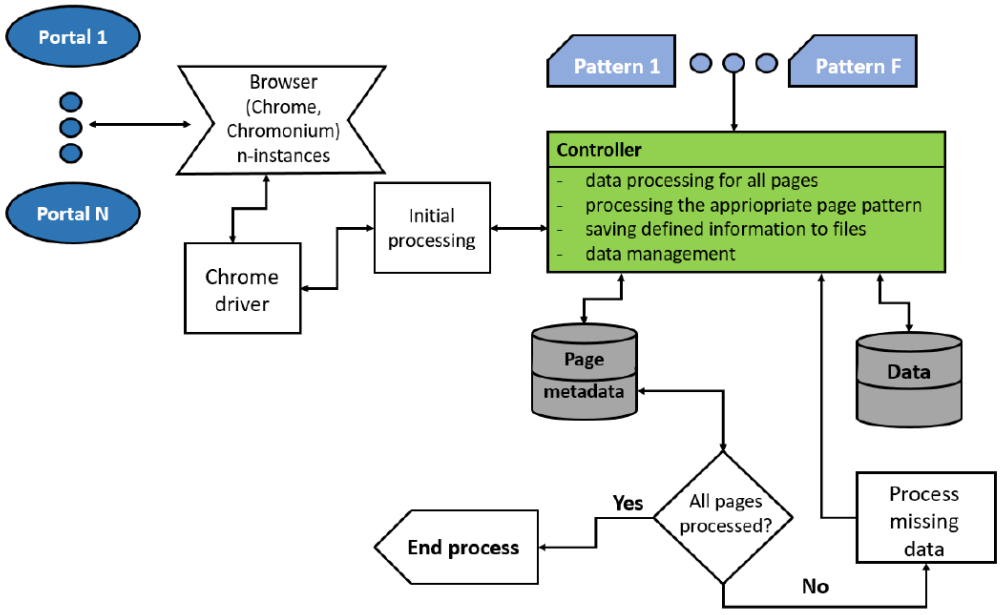
Stage 1: Methods for web scraping to gather the INPUT data, i.e. identification of websites, analysis of the catalogue of websites, analysis of legal aspects. Note that the collected data are not structured and not suitable for immediate analysis.

Stage 2: Data processing steps to PROCESS data, i.e. implementation and saving of extracted data in a temporary database, cleaning and standardization, loading the transformed data into the data warehouse.

Stage 3: Data analysis, final stage to generate OUTPUT data - analysis of the collected data. (Eurostat, 2019a, p. 6)

The web scraping tool was built using primarily Java and Selenium. The use of Selenium is particularly useful in this case, since it is a web automation tool that allows to render HTML that has been dynamically generated with JavaScript or Ajax, and it can also be used to automate certain processes that require human interaction with the website, such as clicking, writing, and submitting forms. A simplified diagram of the architecture of the proposed web scraping tool is shown in Figure 8.

Figure 8. Simplified diagram of the architecture for the web scraping tool



Note. Reprinted from "Deliverable J2: Interim technical report showing the preliminary results and a general description of the methods used" by Eurostat, 2020, *WPJ: Innovative Tourism Statistics*, p. 21.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

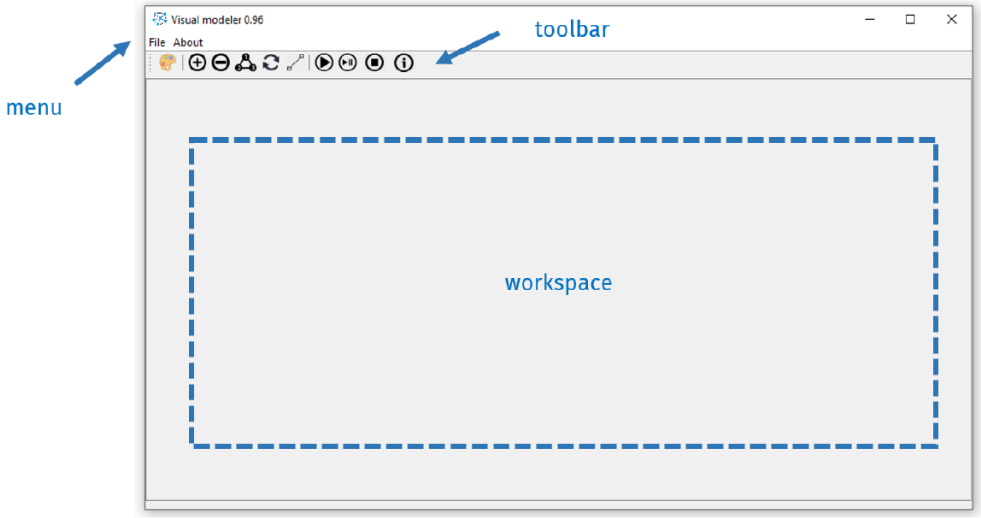
As can be seen, the tool iteratively collects and stores data in the form of HTML files and metadata, with the iteration being repeated until the complete set of data is obtained. In addition, it is possible to define different intervals for each web scraping service. Furthermore, the tool can be executed both in Windows and Linux operating systems, and the optimisation work carried out during the development of the tool contributed to the high degree of efficiency regardless of the platform used.

1.3.3.4 Visual Modeler – Proprietary Tool for Downloading Data

One of the main principles adopted during the design of the web scraper tool was the ability to analyse the entire process quickly and effectively without the need for additional applications. For this purpose, a graphic interface, called Visual Modeler, was developed to easily initialize a web scraping process without specialized computer knowledge (Eurostat, 2020a).

In particular, the graphical interface consists of a visual environment for effortlessly modelling web scraping activities, as shown in Figure 9. To summarise, the process of operating the tool consists of assigning appropriate commands to the individual nodes and entering only the key information required to retrieve the data from the portals. Then, the mechanisms implemented in the tool automatically process the information entered by the person who creates the web scraping process into a set of instructions that perform specific tasks.

Figure 9. Visual Modeler interface



Note. Reprinted from "Deliverable J4: Technical Report" by Eurostat, 2020, *WPJ: Innovative Tourism Statistics*, p. 12.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

In fact, the use of this graphical interface poses the following advantages for the users:

> saving the time needed to start the process, easy and quick possibility to make changes in the web scraping process of the portal, simplified analysis of the correctness of the created process and quick identification of errors, embedded functionalities to provide guidance as proposed by WPC, only basic knowledge of HTML and CSS is required, possibility to write/read processes, possibility to export data to files of different formats, easy extension of the functionality of the tool by adding additional modules, processes do not require compilation [because] they can be run directly ... into the environment, [and the] possibility to work without modifications in Linux, Windows, Mac OS. (Eurostat, 2020c, pp. 23–24)

1.3.3.5 Tourism Integration and Monitoring System (TIMS) Prototype

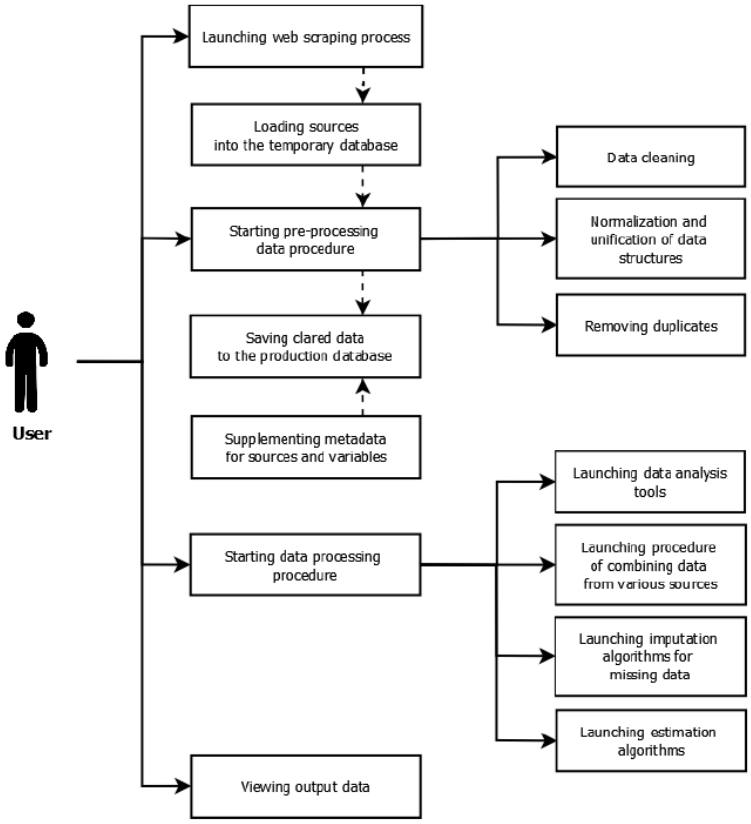
The previously mentioned tools and initiatives serve as the basis for the development of a common Tourism Integration and Monitoring System (TIMS), and as part of the WPJ, a TIMS prototype has been developed.

The tool intends to incorporate into the system databases the newly identified tourism data sources. Furthermore, it aims to support the process of producing statistical tourism data (including experimental data), as well as data from statistical databases (statistical surveys), data provided by external suppliers and data obtained from big data sources (such as web scraping, or social media). Additionally, the system will support several data pre-processing steps, including data cleaning, structural unification, and connection to databases. Finally, the generated output data will be disseminated through existing official statistics networks and through a dedicated API (Eurostat, 2020c). A use case diagram illustrating the possible interactions with the system can be seen in Figure 10.

An overview of the needs of the WPJ partners involved in the ESSnet Big Data II project and an inventory of available big data sources in tourism revealed the need to establish a European-level universal framework tailored to the specific needs of each country. As a result, the system prototype uses a modular structure to easily modify and expand the currently implemented elements. So far, four different functionalities have been fully developed: navigation (moving within and between modules), view (displaying user data), user actions (processing and updating data), and data binding (separating data from the user's view and actions).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Figure 10. TIMS use case diagram



Note. Reprinted from "Deliverable J4: Technical Report" by Eurostat, 2020, *WPJ: Innovative Tourism Statistics*, p. 10.

The system is planned to incorporate support for different data sources, namely statistical databases, data obtained through web scraping, open databases provided by external entities, and social media data. The subsequent elements of the system are related to filling in missing data with the use of machine learning algorithms through cloud-based services (Amazon AWS, Microsoft Azure, Google Cloud, OpenStack, Microsoft Hyper-V, VMWare, or Oracle VM).

Concerning the collection of data, the system will feature a specific module that enables loading data from multiple sources and in different formats, for instance, web scrapped data, relational databases, csv, txt, and xlsx files. The module also allows to directly connect to APIs and databases to collect data based on specific searches or conditions.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

As previously mentioned, the Visual Modeler works as an integral part of the system, allowing the creation of web scraping processes directly within the application itself, which guarantees easier access to new sources of big data by all statistical organizations regardless of their level of programming knowledge.

Furthermore, two of the most powerful modules within the system are the data pre-processing module and the integration module, which include functionalities related to the unification of data types, data cleaning, removing contradictions, and combining information from various sources (Eurostat, 2020c).

It is important to emphasise that missing data can be critical during the data processing procedure, and thus effective imputation methods (techniques for replacing missing data) need to be clearly defined and identified. In view of the fact that the range of imputation methods used in the countries of the EU countries is very different, different methods should be considered, including Mean imputation, Hot Deck Imputation, Cold Deck Imputation, Maximum Likelihood Estimation, Deep Learning, and Multivariate Imputation by Chained Equation (MICE) (Eurostat, 2020c).

Finally, following the data imputation process, further statistical procedures will be available either directly within the system, or as an integration of dedicated software, for instance: “calculation of weights, calibration of weights (linear, raking, logit, truncated), data aggregation (Horvitz-Thompson estimator, Calibration estimator, etc.), estimating precision of results (bootstrapping and resampling), seasonal adjustments (TRAMO-SEATS and X-13ARIMA-SEATS), [and] benchmarking (Denton’s additive and multiplicative approach, etc.)” (Eurostat, 2020c, p. 16). The functional diagram for the TIMS prototype containing the previously described modules can be observed in Figure 11.

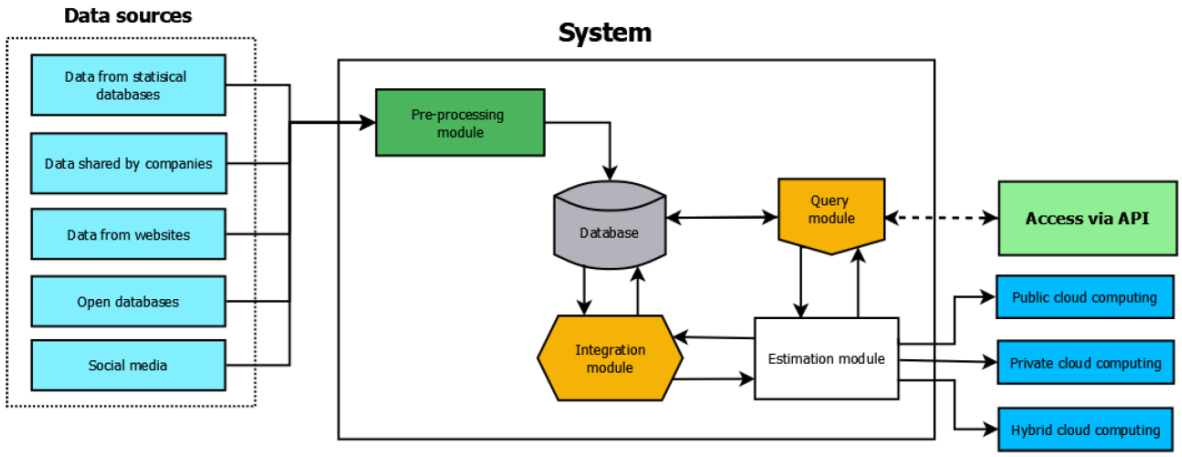
In regard to security concerns, the approach to system security would be similar to the commonly used standards in official statistics. Applicable principles, security regulations, procedures, and responsibilities, including the United Nations Fundamental Principles of Official Statistics, will be fully implemented to ensure an appropriate degree of security within the system.

However, it is necessary to add that a complete system solution will not be fully available during the completion of the project. Nevertheless, the already developed tools

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

and prototypes provide a valuable basis for a common EU framework for the use of big data sources in official tourism statistics.

Figure 11. Functional diagram for the TIMS prototype



Note. Reprinted from "Deliverable J4: Technical Report" by Eurostat, 2020, *WPJ: Innovative Tourism Statistics*, p. 11.

1.3.3.6 Data Linkage

As part of WPJ several data linkage methods were tested for accommodation data collected from online sources. In this case, data linkage consists of matching the data obtained both from survey sources and accommodation portals. Since using simple address data for this purpose is prone to a wide variety of errors, it was proposed to apply a distance-based approach by employing latitude and longitude information to combine data on tourist accommodation establishments from different sources (Eurostat, 2020a).

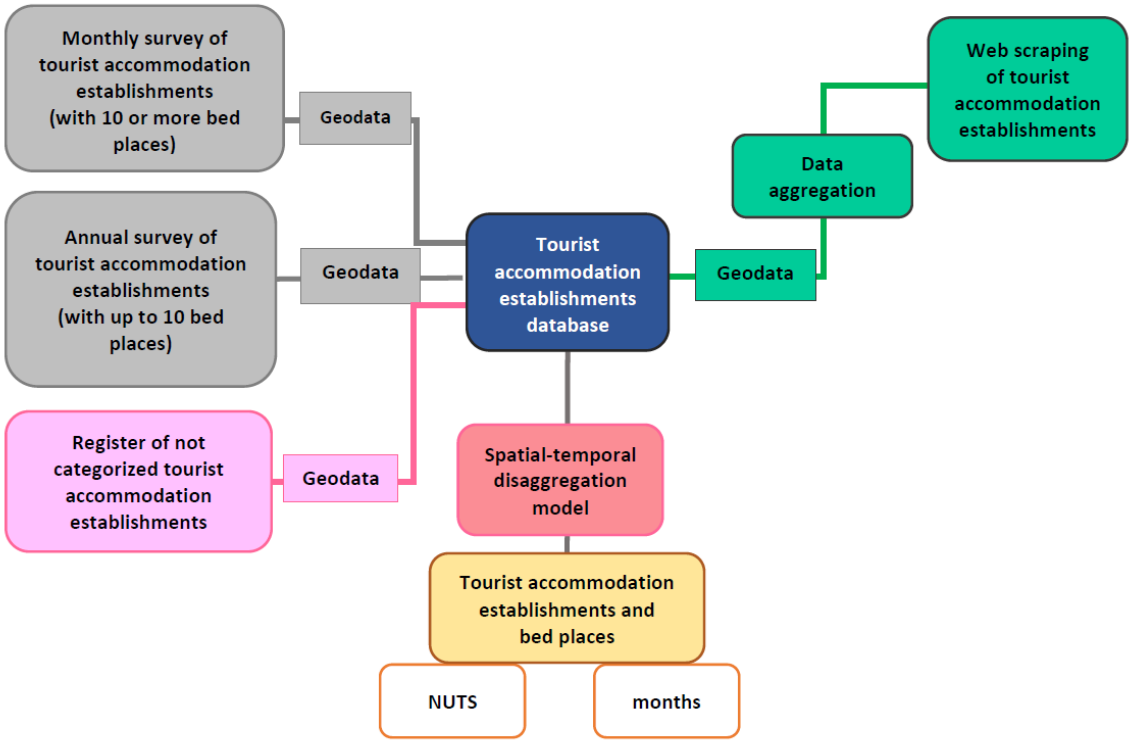
Using this strategy, WPJ has developed a tool for the geolocation of address data from accommodation establishments and web scrapping sources, the tool was built in JavaScript in conjunction with the HERE Maps API. In this sense, if geographic coordinates are available for the survey data as well as for the accommodation portal data, then the distance-based approach for data linkage can be applied. An illustration of a proposed process for data linkage is shown as an example in Figure 12.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

As explained by Eurostat (2020a) the distance-based approach works in the following way:

> Step 1. Derive geographical coordinates for all establishments present in the survey sources as well as in the web scraped data.
>
> Step 2. Calculate the distance between all establishments in the survey sources and all the establishments in the web scraped data.
>
> Step 3. For each establishment in the web scraped data find all the establishments in the survey sources such that the distance between the two does not exceed a specified threshold.
>
> Step 4. For each establishment in the web scraped data, match to the closest one found in the third step. (p. 28)

Figure 12. Example of accommodation data linkage process



Note. Reprinted from "Deliverable J2: Interim technical report showing the preliminary results and a general description of the methods used" by Eurostat, 2020, *WPJ: Innovative Tourism Statistics*, p. 30.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

In an effort to better understand the appropriate linking methods for big data sources in tourism, different case studies were evaluated as part of WPJ. For example, in the case of Netherlands,

> Statistics Netherlands has been using two different tools developed and deployed by Statistics Poland, specifically a web scraper for Hotels.com and a geolocation tool since 2019. The data collected with these tools are meant to keep the tourist accommodation establishments survey information up to date. Pre-processing steps such as data standardization, deduplication, and prioritization are carried out to guarantee comparability across different data sources, as well as the ability to link them to enrich the survey frame. For the most part the (deterministic) linking between Hotels.com data and tourist accommodation establishments is straight forward since it is based on postal code and house number information (Eurostat, 2020b, pp. 11–18).

In summary, the inventory of big data sources related to tourism statistics has shown that around half of them can be used for the implementation of the project objectives. Some of the sources still need to be agreed with the administrators in terms of availability, others need to be further identified as to their quality and usability. Therefore, further work on identifying data sources has been recommended, specially sources related to modern technologies, such as Smart City sources, which will allow to compile statistics for small areas, as well as to provide data on tourist traffic. As part of the project work, a universal tool equipped with a graphic interface has been developed, which allows the web scraping process of online accommodation data to be easily performed. Undoubtedly, a very important stage of the work undertaken in the project has been the development of appropriate methods for combining data from various sources (Eurostat, 2020a).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2 Big Data Initiatives in Tourism Relevant for the MED Region

The following section provides an overview of the main aspects of numerous big data initiatives in tourism relevant for the MED region. Specifically, tourism big data projects from Croatia, France, Greece, Italy, Portugal, Slovenia, and Spain have been examined, as well as EU-wide projects relevant to the MED region. More in-depth information about these particular initiatives, as well as other interesting tourism big data projects outside the MED region can be found in the accompanying Excel file for this report.

Finally, it is important to emphasise that these initiatives represent big data projects mostly from public and academic institutions or resulting from public-private partnerships. Although there is a growing number of private companies offering paid services related to the use of big data in tourism, they were not considered as part of the scope of this research. In general, these companies extract data from websites such as TripAdvisor, Booking, Expedia, Twitter, Instagram, Facebook, Flickr, Google Trends, etc., in order to provide real-time estimates for the tourism sector, and with the use of AI and machine learning techniques they are able to provide assessments and forecasts on particular tourism topics.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.1 EU Institutions

Name of the initiative	Eurostat Experimental Statistics: World Heritage Sites
Main Stakeholder	Eurostat
Location	European Union
Description	Statistics on UNESCO World Heritage Sites (Wikipedia page visits) are taken as an indicator of site popularity or as a measure of 'cultural consumption' of world heritage. The pilot project used monthly page views data for all articles in the 31 Wikipedia available language versions. The data is freely accessible through the Wikimedia Foundation. The pilot was run as part of the Big Data Sandbox, an international joint initiative funded by the Conference of European Statisticians' High-Level Group for the Modernisation of Official Statistics. Besides Eurostat, it included national statistical institutes and other international statistical bodies.
Data	The number of Wikipedia page views and the content of the articles were collected from the different data sources available. As a result, the following statistics were calculated: Top 20 World Heritage Sites in number of page views of related Wikipedia articles (2015), top 5 World Heritage Sites in number of page views of related Wikipedia articles by language (2015).
Link	https://ec.europa.eu/eurostat/web/experimental-statistics/world-heritage-sites
Contact	ESTAT-WIH@ec.europa.eu

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	ESSnet Big Data II: WPJ - Innovative Tourism Statistics
Main Stakeholder	Eurostat
Location	European Union
Description	One of the focal points of the ESSnet Big Data II project is centred on innovative tourism statistics. Specifically, the Work Package J (WPJ) aims to resolve the need for a conceptual framework in tourism big data by creating a smart Tourism Information System. In this way, the system seeks to incorporate innovative statistical techniques to integrate multiple big data, administrative, and statistical sources related to tourism. An overview of the needs of the WPJ partners involved in the ESSnet Big Data II project and an inventory of available big data sources in tourism revealed the need to establish a European-level universal framework tailored to the specific needs of each country. As a result, the system prototype uses a modular structure to easily modify and expand the currently implemented elements. So far, four different functionalities have been fully developed: navigation (moving within and between modules), view (displaying user data), user actions (processing and updating data), and data binding (separating data from the user's view and actions). The system is planned to incorporate support for different data sources, namely statistical databases, data obtained through web scraping, open databases provided by external entities, and social media data[2].
Data	Accommodation data from online booking portals. Flash estimates. Tourist traffic estimations according to travel directions, means of transport, types of accommodation. Prices of accommodation, restaurants, transport, and ticket prices for tourist attractions. Tourism Satellite Accounts. Smart City systems data. Number of people traveling in relation to fairs, conferences, etc. Tourism potential and attractiveness.
Link	https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WPJ_Innovative_tourism_statistics
Contact	Marek Cierpiał-Wolan: M.Cierpial-Wolan@stat.gov.pl

[2] A more in-depth description of the ESSnet Big Data II: WPJ - Innovative Tourism Statistics project can be found in section 1.3.3.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.2 Croatia

Name of the initiative	Dubrovnik Visitors Prediction
Main Stakeholder	City of Dubrovnik
Location	Croatia
Description	Dubrovnik Visitors service provides a realistic estimation of the number of people currently present in the Old Town in the City of Dubrovnik. Another important ICT tool is the Dubrovnik Card Smartphone App (iOS & Android) which works as a digital profile for the most important cultural heritage sites in the city. Dubrovnik Visitors is expected to be used in combination with the Dubrovnik Card app to ease some of the pressure from over-exploited cultural heritage sites. The approach is incorporate the features of the Dubrovnik Visitors system, counting people in the Old City area in real time, with the Dubrovnik Card app. For example, in the event that the number of people exceeds a certain level (4000 people – current suggestion) any user of the Dubrovnik Card Smartphone App would be encouraged to visit other parts of the city that are not usually overcrowded.
Data	The system provides statistics on the estimated number of cruisers and guests, the number of overnight stays, weather details, weather prediction, temperature, precipitation, and additional data from cameras, counters, and sensors. It is planned in the future to also include data from web booking platforms. In this way, the system provides data-based predictions by implementing machine learning algorithms.
Link	http://www.dubrovnik-visitors.hr/prediction
Contact	Nataša Mirić: nmiric@dura.hr

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.3 France

Name of the initiative	DIAMS - Digital Alliance for Marseille Sustainability
Main Stakeholder	Aix-Marseille-Provence Metropolis
Location	France
Description	The DIAMS initiative, which stands for the Digital Alliance for Marseille Sustainability, has four key objectives: to enhance knowledge on air quality and to deliver high quality, reliable big and open data. To encourage the transmission of territorial data and air quality data between local, regional and national channels and ensure their accuracy. To stimulate innovation and leverage the resources of residents and the business sector to co-develop and introduce creative ways to enhance air quality. To provide citizens and decision makers with customised and adaptable knowledge to raise their understanding and engagement.
Data	The DIAMS portal delivers air quality data and resources in the Aix-Marseille-Provence Metropolis supported by numerous contributors. It promotes transactions between various types of entities and organisations that would otherwise be difficult to link together; and offers technical building blocks that are used as a basis on which a large range of innovators may create complementary services and goods. Sensor platform: Sensor collection and assimilation and aggregation procedures. Territorial platform: APIs for platforms collaboration and data sharing, Lean Pollution Inventory, Cloud based conceptual research tools. Innovation portal: access to accessible public data and regulated access to individual data; open innovation network focused on blockchain technologies. Services platform: user-centric applications and community management software.
Link	https://uia-initiative.eu/en/uia-cities/aixmarseille-provence-metropole
Contact	Céline Sales: celine.sales@ampmetropole.fr
Note	Although this project does not specifically deal with tourism data, it may be useful for introducing big data solutions to sustainability issues in the tourism sector.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	DATAtourisme
Main Stakeholder	Directorate General for Enterprises (Direction Générale des Entreprises)
Location	France
Description	DATAtourisme is a national scheme run by the Directorate General for Enterprises, in partnership with the Tourisme & Territoires network, and co-constructed with the networks of French tourist offices and regional tourism committees. It aims to facilitate access to public tourist information data through a national Open Data platform and an API. DATAtourisme device is a data aggregation platform, capable of bringing together data from various local public databases ("tourist information networks and systems"), as well as standardising, and making them available in Open Data via a single access point. The portal collects data from more than 40 different official tourist databases. However, data on visitor numbers are not part of the current scope of the DATAtourisme system.
Data	Festivals and events: cultural events (concerts, exhibitions, festivals, etc.), social events (carnival, traditional festivals, etc.), commercial events (markets, fairs, etc.), sporting events (competitions, demonstrations, etc.). Places: cultural sites (museum, civil building, etc.), natural sites (lake, cliff, viewpoint, etc.), restaurants, accommodation, transport (car park, railway station, public transport station, etc.), shops, practical services (service points for camper vans, bicycle hire, etc.). Products: visits, training courses, tourist activities. Itineraries: tourist routes.
Link	https://www.datatourisme.gouv.fr/
Contact	contact@datatourisme.fr

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Flux Vision Tourisme
Main Stakeholder	Orange
Location	France
Description	Flux Vision Tourisme is a solution developed by Orange, co-constructed with Tourisme & Territoires and used by the majority of the Agencies of Tourist Promotion in France since 2012 (Bouches-Du-Rhône, La Manche, Bretagne, Calvados, Occitanie, Pyrénées-Orientales, Dordogne, etc). This solution enables millions of technical data from the Orange mobile network to be converted into statistical indicators to analyse the use of the territories and the movement of populations. A segmentation (qualification of a mobile phone as "tourist", "resident", etc.) is carried out considering the billing address and the length and frequency of stays. An adjustment is made to move from x mobiles to y people, considering a set of factors such as the rate of mobile equipment and Orange's market share.
Data	Attendance of events without ticketing and with high expected attendance. Tourist mobility by territory. Monthly overnight stays by French tourists. School holiday attendance. Attendance (avg. day) of public holidays. Monthly overnight stays by international tourists. Origin of domestic and international tourists. Overnight stays by destination and time period.
Link	N/A Additional information can be found in: Big Data, Mobilité & Frequentation Touristique (in French).
Contact	Marie Lansonneur: mlansonneur@visitprovence.co

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Big Data & Tourisme
Main Stakeholder	De Vinci Research Center
Location	France
Description	The "Big Data & Tourisme" project is based on the creation of a web-based data mining tool to monitor the tourism impact of the development of Bordeaux. The aim is to create an interface for monitoring and analysing tourist practices by providing big data from various social networks in order to model and understand tourist behaviour in Bordeaux and its region and assess the impact of recent changes in the area (Cité du Vin, LGV...). Identify and quantify in relative terms the regional tourist flows entering and leaving Bordeaux on the basis of the traces left in the other tourist centres in the region. Identify synergies, if any, between different tourist centres in the Bordeaux metropolis and also between tourist basins or regional or even extra-regional tourist centres.
Data	Effects on tourism of the launch of "La Cité du Vin" and the introduction of the LGV (high-speed rail). Online tourist reviews, geotagged and geo-localized photos.
Link	N/A Additional information can be found in: Villes Intelligentes (in French).
Contact	Arnaud Beyssen: arnaud.beyssen@caissedesdepots.fr

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.4 Greece

Name of the initiative	InnoXenia: Innovation in Tourism in the Adriatic-Ionian Macroregion
Main Stakeholder	Region of Western Greece
Location	Greece
Description	The aim of the InnoXenia project was to establish an innovative tourism platform (Tourism Innovation Observatory and Tourism Innovation Decision Support System) that would enable stakeholder networking in the destination, knowledge transfer and better elaboration of tourism strategies, policies and investments in the Adriatic – Ionian area.
Data	Tourism Destination: Total tourism expenditure over GDP, direct jobs in tourism, quality of life index. Tourism Industry: Tourist accommodation establishments, museums, archaeological sites, net occupancy of beds, average length of stay, capacity ratio, tourists per capita, tourism impact on population, tourism impact on destination. Tourism Flows: Domestic arrivals, arrivals for personal reasons, business arrivals. Accommodation: Average price per month, average rate. ETIS: Tourism enterprises in the destination using a voluntary certification for environmental and corporate social responsibility measures.
Link	https://www.innoxenia.eu/index.html
Contact	Kosstantinos Tzamaloukas: ktzam@ilia.pde.gov.gr

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.5 Italy

Name of the initiative	HERIT-DATA
Main Stakeholder	Tuscany Region (Regione Toscana)
Location	Italy
Description	HERIT-DATA: *Innovative Solutions to Better Manage Tourism Flows Impact on Cultural and Natural Heritage Sites through Technologies and Big Data* is an Interreg MED project with the goal of reducing the effect of tourism on cultural and natural heritage sites through the use of technologies and big data in particular MED regions (Croatia, France, Greece, Italy, Spain and Bosnia Herzegovina). One of the key objectives of HERIT-DATA is to develop, test and launch a Tourist Flow Management Platform based on Snap4City (scalable smart analytic application builder for sentient cities and IoT). The platform is expected to collect and analyse data from various sources: social media, sensors, ports, PAX counters for counting museum and event visitors, transportation, and other platforms. So far, a demo for a Twitter Vigilance Dashboard has been launched and data are expected to be integrated according to geographical information.
Data	Average number of people every 15 minutes based on Wi-Fi sensors, number of people in transit between two points, average time spent in sensor location, number of people every minute based on camera position data. Twitter indicators: Number of hashtags, users, keywords, tweets, retweets and citations. Most frequent terms. Sentiment Analysis (positive / negative) for adjectives, nouns, and verbs in English and Italian. NLP extraction of adjectives, nouns, verbs, hashtags, and usernames in English and Italian.
Link	https://herit-data.interreg-med.eu/ https://rttvhd.snap4city.org/?p=chart_utente
Contact	General contact: heritdata@regione.toscana.it Sergio Papiani: sergio.papiani@regione.toscana.it Davide Bruno: davide.bruno@regione.toscana.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	ShapeTourism: New Shape and Drive for the Tourism Sector
Main Stakeholder	Ca' Foscari University of Venice
Location	Italy
Description	ShapeTourism is an Interreg MED project created with the purpose of developing a system of tools called Smart Tourism Data System, developed to explain and monitor sustainable tourism within the MED region, along with a support system for data collection and analysis. Specifically, the following tools were developed: ShapeTourism Observatory, ShapeTourism Survey, ShapeTourism Carrying Capacity Scenarios Simulator, and ShapeTourism Clusters Maps. The ShapeTourism Observatory integrates data from traditional statistical sources with big data sources in relation to the NUTS 1, 2, and 3 levels. In particular the Observatory monitors four different categories of information: Reputation, Attractiveness, Competitiveness, and Sustainability.
Data	Reputation: Reputation data are presented in relation to attractions, rentals, restaurants, and hotels. The main variables are average rating, weighted rating, average number of reviews, total reviews, average price, etc. Attractiveness: Monuments and other tourist sights, indexed, population density (10,000 inhab. per km^2), number of beds in hotels and similar establishments per inhabitant, accessibility, etc. Competitiveness: Travel and Tourism Competitiveness index (TTCI) calculated on the basis of business environment, safety and security, health and hygiene, human resources and labour market, ICT readiness, prioritization of travel & tourism, international openness, price competitiveness, environmental sustainability, etc. Sustainability: GDP per inhabitant pps, GDP per inhabitant pps, (log), arrivals in hotels and similar establishments, arrivals in hotels and similar establishments, regional sustainability index (1-7), etc.
Link	https://shapetourism.interreg-med.eu/
Contact	Jan van der Borg: vdborg@unive.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Smart Destination
Main Stakeholder	Tuscany Region (Regione Toscana)
Location	Italy
Description	Smart Destination aims to develop and test an intelligent technological system in which businesses, tourism associations, cross-border operators, public administrations, land management bodies and operators in the tourism industry can collaboratively integrate data and information on the services and products offered in order to respond competitively to tourists' requests. Specifically, Smart Destinations intends to design an open and standardized system that is able to aggregate fragmented tourist information between numerous data platforms, enable unified access for the different management systems of public and private actors in the tourism field, build an integrated offer of personalised tourism services, and develop advanced big data analysis functionalities (sentiment analysis, demand forecasts, etc.).
Data	Statistical information and tourist movements using the "Smart Tour" app data. The Smart Destination monitoring system (“Smart Monit”) plans to integrate data from a range of different sources: A database containing information related to parking, transportation, mobility and events. A database containing information regarding hotels, restaurants, and museums. Data from regional Tourist Destination Observatories (OTD) regarding tourist arrivals, hotels, number of bed places, and accessibility. Data produced by the app "Smart Tour", specifically concerning art and museums, food and beverages. hotels and reservations, live streaming, content sharing, transport, GPS services, booking services, traffic info, infrastructure, events, itinerary planning. Social media and online news data for sentiment analysis.
Link	http://interreg-maritime.eu/it/web/smartdestination
Contact	Carolina Gentili: carolina.gentili@regione.toscana.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Turismo Big Data
Main Stakeholder	Unioncamere
Location	Italy
Description	The platform Turismo Big Data, is a municipal level data system, extended to the whole Italian territory, which offers an overview of the tourism sector, containing information coming from multiple web sources. The platform offers an overview of the tourism sector, using a series of statistical variables relating to: tourism in the strict sense of the term, i.e. those businesses that constitute the core business of the Italian tourist offer such as accommodation facilities (hotels and complementary), restaurants and other catering establishments in the area, bathing establishments, tourist brokerage agencies, etc.; transport and tourist mobility infrastructure such as airports, rail transport, taxi transport, car rental with driver, land transport of passengers in urban and suburban areas, etc.; cultural resources such as museums, monuments and archaeological sites, but also cultural routes, certified Made in Italy food and wine, and natural resources; the companies that make up the so-called extended chain of tourist hospitality such as leisure time entertainment, sports, recreational activities, theatres, entertainment venues such as discos or dance halls, etc.
Data	Indicators and trends: Total tourism enterprises, total number of employees, tourism enterprises in the national total, tourists' expenditure (accommodation facilities, private homes), social networks (companies present, engagement capacity, private accommodation), tourism flows in accommodation enterprises, resources (museums and cultural sites, national parks, blue flags, certified products, etc.), main holiday motivations, average spending, communication channels that influence the choice of the stay, and activities carried out during the holiday.
Link	https://turismobigdata.isnart.it/
Contact	Unioncamere: unioncamere@cert.legalmail.it ISNART: info@isnart.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Smart Tourism and Big Data
Main Stakeholder	Piedmont Region (Regione Piemonte)
Location	Italy
Description	This exploratory project analysed the tourism sector in Piedmont during the period May-October 2015 using anonymous and aggregated data from the Vodafone Italia network. The project focused on four main aspects: Analysis of the context, to highlight the main indicators of Italian and foreign tourism in the region. Analysis of the segments, to identify how tourists divide the territory on the basis of their behaviour. Analysis of the dynamics, to highlight the connections and flows between the places in the region. And analysis of experiences, to verify the centrality of places in the visitor's overall experience. This study is part of a series of initiatives from the Piedmont Region and the Regional Tourism Observatory to assess the feasibility of adopting big data tools. As a result, the Regional Tourism Observatory has started to adopt tools and services based on mobile phone data that allow to trace the tourist presence in the territory, but these tools are not publicly available.
Data	Presence of foreign visitors in Piedmont. Presence of Italians in Piedmont. Foreign visitors in Piedmont and in the main cities, classified by country. Italian visitors in Piedmont, classified by region and municipalities of origin. Foreign visitors in local tourist agencies, classified by country. Italian visitors in local tourist agencies, classified by municipalities. Temporal trend of foreigners' visits in Piedmont. Temporal trend of foreigners' visits in the cities of Piedmont. Temporal trend of Italians' visits. Weekly trends. Permanence and repetitiveness of visits. Connections between the cities (number of visitors who have visited both locations). Intensity of connections between municipalities in Piedmont. Segmentation of the territory by tourists. Composition of tourists in the segmentation of the territory. Territorial preferences.
Link	N/A Additional information can be found in: Smart Tourism e Big Data (in Italian).
Contact	Visit Piemonte: dmopiemonte@legalmail.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	SMOOTH Venice: Spatial Modellization of Territorial Heritage
Main Stakeholder	Chamber of Commerce of Venice Rovigo Delta Lagunare
Location	Italy
Description	The project aims at correlating, according to an innovative approach based on three-dimensional and geo-referenced spatial analysis, strategic development indicators that consider territorial potential. But above all the dynamics that have affected the area over the last ten years, determining shifts in economic values and the location of the population and businesses.
Data	Size and location of resident population and enterprises (analysis on master data, ISTAT censuses and Chamber of Commerce register data); land consumption and land use; this data band serves to represent and describe the metropolitan territory in the size of land use (Corine Land Cover database) and land consumption (Veneto Region database); real estate values and transactions; this database serves to represent the value of the land and especially the value of the residential and non-residential components, both in monetary terms and in terms of buying and selling dynamics, through the consultation and purchase of the OMI-Agenzia delle Entrate database, which will be used to define the variables that allow to qualify or de-qualify a territory according to the settlement and infrastructural changes linked to the other databases, according to elements of territorial and statistical correlation; infrastructural endowments and mobility systems; this database will use the georeferencing system of data towards the Google Maps system and will also include a verification of the availability of dynamic flow indicators, in particular on Waze databases, in order to build a system of indicators capable of linking static information on the territory to dynamic information on land use and infrastructure.
Link	http://quantitas.it/data/smooth-venice.it/httpdocs/
Contact	Quantitas: info@quantitas.it
Note	Although this project does not specifically deal with tourism data, it may be useful for introducing big data solutions to territorial heritage issues in the tourism sector.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	DARE: Digital Environment for Collaborative Alliances to Regenerate Urban Ecosystems in Middle-Sized Cities
Main Stakeholder	Municipality of Ravenna
Location	Italy
Description	DARE is a three-year project under the EU initiative Urban Innovative Action (UIA). DARE aims to demonstrate the effectiveness of a citizen-centred, digital-based governance approach aimed at facilitating, supporting and accelerating the implementation and evaluation of the regeneration process of an entire neighbourhood, that of the Ravenna Dock. The project proposes the creation of a three-level digital environment, consisting of a data management platform (data level), a content management system (editorial level) and a so-called ViR-Virtual Realm (presentation level), intended as the enabling technology needed to activate urban actors. The digital infrastructure will enable the collection, management and creation of data relating to the dock (e.g. vehicle traffic, economic data on the area's activities, environmental and population data, etc.), thus making available real-time snapshots of the environmental, social and economic situation in the area.
Data	The digital infrastructure will enable the collection, management and creation of data relating to the dock (e.g. vehicle traffic, economic data on the area's activities, environmental and population data, etc.), and quality of life indicators, thus making available real-time snapshots of the environmental, social and economic situation in the area.
Link	https://www.dare-ravenna.eu/
Contact	Municipality of Ravenna: UPE@COMUNE.RAVENNA.IT

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Optimising Tourism in Tuscany
Main Stakeholder	Toscana Promozione Turistica
Location	Italy
Description	In collaboration with Toscana Promozione Turistica, Data Science for Good has worked to research the travel habits of tourists in Tuscany, to recognise the different types of tourists visiting Tuscany: “city-hoppers” who focus their trips in major cities; “coast-lovers” who spend much of their time along the coast; “adventurers” who frequent both the coast and inland cities and attractions; and the “countrysiders”. Through extracting data features and using machine learning algorithms to group individuals, it was possible to identify and better understand the people visiting the area. The goal of the project was to assist local authorities in identifying and evaluating tourism beyond conventional surveys and figures so that they can explore and develop sustainable tourism strategies in the region. Vodafone Italy provided anonymized data that was used to classify trends of tourism in time and space.
Data	Location clustering analysis by season, country of origin of visitors, tourist mobility by typology (city hoppers, coast lovers, explorers, countrysiders), cluster analysis of tourist trajectories by country of origin and season.
Link	http://dssg-eu.org/tuscany/index.html
Contact	Data Science for Good: info@datascienceforsocialgood.org

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Multi-Access Digital Platform on Visitor Tourism Data
Main Stakeholder	Italian National Tourism Board (ENIT)
Location	Italy
Description	Data related to components of the tourism offer and its performance, rates applied on OTAs and related occupancy data, visitors, markets, types of travel, channels used, events by type and forecasts of impact in number of visitors.
Data	Occupancy rate of tourism SME's, visitor characteristics and events by period, areas and sub-areas, canals, clusters, origins, typology and characterized by product areas (seaside, mountain, inland areas, cities, spas, lake, commercial) and reasons (e.g. cultural, food and wine, sports, etc.).
Link	N/A
Contact	Elena Di Raco: elena.diraco@enit.it

Name of the initiative	Online Platform for Monitoring Airport Traffic to and from Italy
Main Stakeholder	Italian National Tourism Board (ENIT)
Location	Italy
Description	Data related to airport arrivals in Italian airports for airport of departure from USA, Germany, UK, Spain, France, China, Russia, Netherlands, etc.
Data	Number of arrivals of foreign tourists at Italian airports on the basis of reservations collected in real time by GDS systems by airport of origin, by country of origin, by destination airport, by travel class, by number of passengers per booking, by booking channel, and passenger profile.
Link	N/A
Contact	Elena Di Raco: elena.diraco@enit.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Monitoring and Listening to the Web and Social Media through Social Analytics and Social Listening Services
Main Stakeholder	Italian National Tourism Board (ENIT)
Location	Italy
Description	Web and social media monitoring and listening tool for social analytics and social listening services.
Data	Analysis of online web and social media data. Tracking of traffic volumes and identification of networks and digital environments where the brand or product is discussed. Verification of the degree of engagement generated. Social analytics of social accounts connected to ENIT (head office and foreign offices), Italia.it, Italian Regions, Ministries, consular entities and project partners. The social accounts to be monitored must concern all the following platforms: Facebook, Instagram, Twitter and YouTube. Monitoring of the following types of sources: periodical newspapers, blogs, forums, generic websites. Comparison of social analytics with the accounts of Italy's main competitors also on the basis of socio-demographic, geographical and temporal parameters. Social listening: monitor social channels by creating queries related to topics that can range from listening to a brand (location), a service, a communication campaign, a competitor, etc.
Link	N/A
Contact	Elena Di Raco: elena.diraco@enit.it

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.6 Portugal

Name of the initiative	TravelBI
Main Stakeholder	Portugal Tourism Board (Turismo de Portugal)
Location	Portugal
Description	TravelBI is a data portal developed by the Portugal Tourism Board. The portal contains statistical data and research publications regarding Accommodation, Cities and Urban Tourism, Consumer Behaviour, Cultural Tourism, Employment in Tourism, Events, International Tourism, Market Trends, Medical Tourism, MICE, Sustainability, Tour Operators, and Training. Data are compiled from several official sources such as Turismo de Portugal, INE – Statistics Portugal, ABAE – European Blue Flag Association, Bank of Portugal, DGEC - General Directorate for Energy and Geology, and ICCA - International Congress and Convention Association.
Data	Water in Bathing Areas with Good/Excellent Quality, Percentage of Seasonal Jobs, Accommodation Estab. Accessible for Guests with Special Needs, Environmental Spending per Resident, Tourism Employment by Education Level, Tourism Density, Energy Consumption and Emissions in Tourism, Tourism Employment by Gender, Tourism Intensity, Accommodation Estab. with an Environmental Certification, Accommodation Estab. Open Year Round, Number of Beds Available per 1000 Residents, Accommodation Estab. with Objectives to Reduce Water Consumption, Average Length of Stay, Percentage of Returning Tourists to Portugal, Accommodation Estab. with Objectives to Reduce Energy Consumption, Number of Overnight Stays, Accommodation Estab. that Optimize Water Consumption, Tourism Receipts, Accommodation Estab. that Optimize Energy Consumption, Accommodation Estab. that use Local Suppliers, Accommodation Estab. Separating Different Types of Waste, Average Spending by Same-day Visitors, Accommodation Estab. that Provide Training on Sustainable Practices, Average Spending by Tourist, Urban Solid Waste Attributed to Tourism, etc.
Link	https://travelbi.turismodeportugal.pt/
Contact	Portugal Tourism Board: conhecimento@turismodeportugal.pt

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Tourism Information Portal (Portal de Informação Turística)
Main Stakeholder	NOS
Location	Portugal
Description	The project is composed of a pioneering platform to gather statistical information on tourism in Portugal. With the support of Turismo de Portugal, the project provides access to relevant information on the presence of foreign tourists and the forecast of demand for Portugal as a tourist destination.
Data	Tourist Pressure: Ratio between the number of different tourists in the municipality and its resident population, in one month. Standardised indicator for a scale from 1 (minimum) to 6 (maximum). Tourist Density: Number of distinct tourists per km2 of area of the municipality, in one month. Standard indicator for a scale from 1 (minimum) to 6 (maximum). Tourist Diversity: Number of different countries of origin in each municipality, in one month. Linguistic Diversity: Number of different official languages (of the countries of origin) present in each municipality in a month. Currency Diversity: Number of different currencies (from countries of origin) present in a month in each municipality. Weekenders: Ratio of the average number of tourists per day at the weekend to the week in each municipality in a month. Indicator with a value above 100 means more tourists on weekends than on weekends. Attraction Lunch: Percentage of the number of tourists at lunchtime compared to the total number in each municipality in a month. Dinner attraction: Percentage of tourists at dinner time compared to the total number in each municipality in a month. Night Retention: Percentage of the number of tourists at night compared to the total number in each municipality in a month. Digital Sophistication: Composite indicator, standardised on a scale from 0 to 100, of the possession of smartphones, intensity in the use of data, among others.
Link	https://corporateanalytics.nos.pt/index.html
Contact	NOS: corporate@nos.pt

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

2.7 Slovenia

Name of the initiative	Tourism 4.0
Main Stakeholder	Arctur
Location	Slovenia
Description	The key objective of the project is to create a shared Tourism 4.0 network, a transformative technical approach centred on confidence and the sharing of data between all tourism stakeholders to produce a new generation of tourist apps, facilities and processes while preserving the privacy of users. As part of the Tourism 4.0 project, Slovenia aims to develop a tool called FLOWS. FLOWS will allow advanced analyses and predictions of tourism movements based on anonymised data from a broad range of sources. The platform will be built on state-of-the-art high-performance computing infrastructures supported by AI, blockchain and high-performance data analytics.
Data	Tourism flows, travellers, day-trippers, guests by mobility patterns, number of stops, seasonal deviations, entrances/exit to destination, movement inside destination, etc. Data will be shown at the chosen time period (year, month, week, day), based on historical records, and weighted by relevant criteria (weekend, temperature, national or holidays in other countries, etc.).
Link	https://tourism4-0.org/flows/
Contact	https://tourism4-0.org/contact/

Project co-financed by the European Regional Development Fund

2.8 Spain

Name of the initiative	Measurement of the Number of Tourist Dwellings in Spain and their Capacity
Main Stakeholder	National Statistics Institute (INE)
Location	Spain
Description	This experimental statistics project carried out by the INE aims to complete the national statistics on tourist accommodation (hotel establishments, tourist flats, campsites, rural tourism accommodation and hostels) with data obtained through the webscraping of online tourist accommodation platforms. Traditionally, these statistics contain information from a variety of occupancy surveys. However, to date there is not enough information available on tourist accommodation to enable the impact on the tourist sector to be analysed and the quality of the statistical data to be improved. For this reason, data on tourist dwellings were extracted from various online booking platforms and integrated on the basis of existing tourist housing directories, their geographical distance, and their licence.
Data	Data on tourist dwellings were extracted from various booking platforms and integrated on the basis of existing tourist housing directories, their geographical distance, and their licence. Both the number of tourist homes and their capacity will be published. The geographical breakdowns used will be: Autonomous Community, province, municipality and tourist areas. Specifically the following data were collected: Accommodation identifier, Name of accommodation, Location, Capacity, Rating, Type of accommodation, Sub-type of accommodation, Name of host, Licence, Description of accommodation, Neighbourhood and host, Number of comments, Address, Number of bedrooms, Number of beds, Number of bathrooms, Size, Internet and parking services, Availability of swimming pool, Whether smoking and pets are allowed, and Company managing the accommodation.
Link	https://ine.es/en/experimental/experimental_en.htm
Contact	https://ine.es/infoine/?L=1

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Distribution of the Expenditure made by Foreign Visitors on Visits to Spain
Main Stakeholder	National Statistics Institute (INE)
Location	Spain
Description	The project is based on incorporating information from auxiliary data sources to improve the statistical operations on tourism carried out by the INE. One of the sources of additional information which has been analysed is that provided by the banking transactions carried out through cards by visitors on their trips or excursions. These banking transactions include card transactions carried out in person, i.e. payments made through POS terminals, as well as cash withdrawals from cash dispensers. Taking advantage of this additional information and focusing on the Tourism Expenditure Survey (EGATUR), an experimental statistic has been carried out which has made it possible to determine how much foreign visitors spend on their trips and excursions to Spain according to their country of residence, and in which autonomous communities they spend it, considering all the communities visited and not only the one assigned as the main destination.
Data	Percentage distribution of annual expenditure in actual destination within each Autonomous Community of destination by country of residence of the visitors. Percentage distribution of the annual expenditure made in real destination by each country of residence of the visitors, according to Autonomous Community of destination. Average daily expenditure in actual destination within each Autonomous Community of destination by country of residence of the visitors. Average expenditure per visitor in actual destination within each Autonomous Community of destination by country of residence of the visitors. Average daily expenditure in actual destination within each country of residence of the visitors, according to Autonomous Community of destination. Average expenditure per visitor in actual destination for each country of residence of the visitors, according to Autonomous Community of destination. Percentage by nationalities of card payments with respect to the total expenditure of EGATUR.
Link	https://ine.es/en/experimental/experimental_en.htm
Contact	https://ine.es/infoine/?L=1

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Measurement of National and Inbound Tourism from the Position of Cell Phones
Main Stakeholder	National Statistics Institute (INE)
Location	Spain
Description	The Sub-directorate of Tourism Statistics and Science and Technology of the INE, in collaboration with the three largest mobile network operators in Spain, is currently carrying out a pilot study for the use of aggregate mobile telephone data to determine the movements of resident and foreign tourists and excursionists, breaking down the information by Autonomous Communities, provinces and municipalities. And in the same way, the country or countries through which national resident tourists travel when they go abroad. The purpose of the project is to improve existing traditional statistics collected through the Survey on Tourism of Residents, and Tourist Movements in Borders.
Data	The variables under study would be the trips, overnight stays, and excursions of national and foreign tourists. Specifically, the aim of the project is to find out in an aggregate way, through mobile phone signalling, by means of active and passive events captured by telephone antennas, the origin of foreign tourists visiting Autonomous Communities, provinces and municipalities; the patterns of tourist behaviour (movements) associated with each nationality; the origin and destination of resident tourists and excursionists visiting the different Autonomous Communities, provinces and municipalities; and the origin and country or countries visited by resident tourists and excursionists on their trips abroad.
Link	https://ine.es/en/experimental/experimental_en.htm
Contact	https://ine.es/infoine/?L=1

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Tourism of Tomorrow Lab
Main Stakeholder	Regional Government of Andalusia (Junta de Andalucía)
Location	Spain
Description	The Tourism of Tomorrow Lab (ToT Lab) is a current initiative of the Digitalisation and Safety for Tourism S3 platform. The ToT Lab will produce business insight and new information that will be used by its participants at regional level to provide capacity development, training and services in the tourism field. Information produced by the ToT will provide input to regional small and medium-sized enterprises in terms of technology growth and universities in terms of training.
Data	The Lab will be the EU centre gathering research capability, expertise, and data on tourism mobility, preferences, demand, etc. No specific indicators have been specified yet.
Link	https://s3platform.jrc.ec.europa.eu/tourism
Contact	JRC S3P Industry: JRC-S3P-INDUSTRY@ec.europa.eu Kristian Sievers: kristian.sievers@lapinliitto.fi Ana Moniche: amoniche@andalucia.org

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	SmartData
Main Stakeholder	Turismo de Andalucía
Location	Spain
Description	SmartData is a tourism data application with several objectives: To make available to the Andalusian business sector and tourist destinations sources of data of tourist interest that will help in the decision-making process. To make the data accessible, in an understandable way, ready to be used regardless of technical training and technological infrastructure. To complement official data, with non-official data sources, which are of special interest to the sector. Promote the use of data in the tourism sector. Developed with Pentaho.
Data	Tourist accommodation places: Hotels, Hotel-Apartments, Pensions, Hostels, Rural Houses, Hostels, Tourist Apartments, Campsites, Tourist Rural Accommodation Housing, Housing for Tourist Purposes. Number of tourists. Number of overnight stays. Length of stay. Airbnb data: number of stays by city, number of available tourist structures by type. Passenger and aircraft arrivals by airport and reference period. Passengers disembarking in the reference period. Cruise ships that dock in the reference period. Movement of passengers by bus to Andalusia by nationality. Key destination-related terms from Google Trends. Sentiment analysis by destinations and topics.
Link	https://smartdata.andalucia.org/
Contact	https://smartdata.andalucia.org/contacto/

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Tourism Data of the Valencian Community
Main Stakeholder	Turisme Comunitat Valenciana
Location	Spain
Description	Tourism Data of the Valencian Community is an interactive report published weekly by the Valencian Institute of Tourism Technology (Invat-tur) to analyse the tourism situation. The report, addressed to the Valencian tourist sector, assesses the current scenario with data on mobility, hotel offer, reservations or flights, among others.
Data	Border openings and mobility, Hotel reservations in the and Hotel offer. Information on trends in Google for travel to the Region of Valencia, shown by province, and the air capacity of the airports of Valencia and Alicante. The data from the Travelgate platform is also analysed to find out about trends in hotel bookings.
Link	https://infogram.com/1p3egzr056zqnxi0r2g6j1ypenadrqvknd6?live
Contact	invattur@gva.es

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Smart Beaches
Main Stakeholder	Turisme Comunitat Valenciana
Location	Spain
Description	This project is framed in the context of the model of Intelligent Tourist Destinations of the Valencian Community (DTI-CV) and pursues the objective of analysing, conceptualising and designing the services and technological tools that a beach or other natural space should have in order to approach the model of intelligent planning and management. There are currently three pilot projects in progress on the beaches of Benidorm, Gandía, and Benicàssim. A Smart Beach is a resource where technology and innovation are applied, adapting it to its specific characteristics to make it a more user-friendly space oriented to satisfy the needs of the tourist, with the aim of improving the visitor experience through more efficient management. Each source provides the data needed to maintain an indicator. Thus, for example, data on beach occupancy provided by a Wi-Fi system will increase in accuracy if crossed with video image analysis of the same area. These relationships provide a new scenario of multiple publication, being able to send interpreted data to the user, publish in Open Data form for its reuse by third parties, such as service companies, or show the information not only on websites and apps but also on transport or security platforms and information panels located in the most interesting or busiest points.
Data	Sensors: UVA, temperature, mobilisation, sediments. Video surveillance: image analysis, monitoring, occupation. Wi-Fi: Occupancy density. Smart buoys: Proximity, boat surveillance, sea conditions, access restrictions, presence of jellyfish, water quality. Bathymetry: Monitoring of the seabed, location of species.
Link	https://www.invattur.es/playas-inteligentes/
Contact	invattur@gva.es

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	DATATUR4CV
Main Stakeholder	Turisme Comunitat Valenciana
Location	Spain
Description	DATATOUR4CV is a tourism intelligence and knowledge transfer system aimed at the analysis, study and dissemination of behaviour and trends in tourism demand with the aim of providing added value to Valencian tourism agents in their process of improving quality and increasing their competitiveness. Objectives: a) To establish a statistical information system that guarantees the reliability of data obtained and their permanent updating. b) To prepare information of a micro and macroeconomic nature on supply and demand in the tourism sector of the Valencian Community. c) To determine the natural evolution of the tourism market. d) To analyse the present and future situation of the national issuing markets and international, specific interests, consumption patterns of different travellers, their technological, social, environmental or any other needs. e) To identify the needs for immediate action, the trends that will allow to anticipate changes and lead the coordination to adapt the tourist offer of the Valencian Community to the real needs of the current tourism market. f) To prepare reports on the tourism situation in the Valencian Community by analysing the evolution of indicators that provide insight into the model.
Data	Specific variables and indicators have not been defined yet. DATATUR4CV plans to use Big Data, Machine Learning and Business Intelligence cloud services to capture, integrate, process and analyse different data sources in order to visualise reports and interactive dashboards that will enable it to become a decision-making support tool for stakeholders in the Valencia tourism sector. To this end, DATATUR4CV will require the selection and acquisition of data through information sources related to tourism indicators and tourist behaviour, as well as the selection and configuration of services in the cloud (SaaS) to process information.
Link	https://www.invattur.es/smart-data-office/
Contact	invattur@gva.es

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	IoT and Big Data in Action: Sagrada Familia
Main Stakeholder	City Council of Barcelona (Ajuntament de Barcelona)
Location	Spain
Description	This project focuses on how IoT and Big Data technologies can improve all management, decision-making and planning activities carried out by local authorities in relation to tourism, in particular, the project analysed the mobility patterns of visitors in the areas of interest of the Sagrada Familia basilica, using MNO and sensor data.
Data	During the first phase of the project, the basic data of 15 million Orange customers was analysed, which allowed, among other things, the identification of: areas of great tourist interest, general profile of visitors, concentrations of visitors per area, and the busiest routes between the neighbourhoods of Barcelona. Mobile phone data was used to measure, detect and quantify people in the city's areas of interest. They have complemented the demographic and tourist information obtained through the surveys and have made it possible to define the mobility patterns between the districts. In order to analyse the profile of the visitors, their behaviour and mobility patterns in the area, other than the IoT elements, were combined and deployed at street level: 9 Wi-Fi sensors - to analyse the flow of visitors and how they move around the area, 1 GSM sensor - to obtain the nationality of the visitor, and 3 3D sensors - to count people entering and leaving the different metro stations
Link	https://d-lab.tech/es/proyecto-1/
Contact	info@mobileworldcapital.com

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Motriz: Regional Tourist Information System
Main Stakeholder	Institute of Tourism in the Region of Murcia (ITREM)
Location	Spain
Description	ITREM has developed Motriz, with the objective of responding to the information needs expressed by the sector to guarantee better information of the environment and the current business models in tourism. In the same way, Motriz allows having predictive models to anticipate changes and new needs of the current tourist and to obtain indicators that measure the competitiveness of the tourist companies, as well as the efficiency of the promotion and marketing actions of the destination. Through Motriz, the sector is provided with a tool that allows the collection, measurement and analysis of tourism data in a simple way for strategic decision making. It also offers the possibility of having, in a single application, all the information that could be necessary for a general knowledge of tourism management in the Region of Murcia.
Data	National tourism and foreign tourism. Occupation in collective accommodation. Airport passengers. Tourist expenditure. INE occupancy surveys for hotels, flats, campsites and rural accommodation in the Region of Murcia. Tourism offer. Origin of users. Reason for visiting the municipality. Type of user. User's age. Information requested. Online prices: these are the prices offered by the hotels in the leading online booking platforms in Spain and Europe. Online availability of accommodation places. Online reputation: Users' evaluations of the aspects of comfort, value, location, cleanliness, services and staff. In the future, it is planned to integrate new data sources such as MNO, tourist spending through credit cards, air traffic, etc.
Link	N/A Additional information can be found in: Sistema de Inteligencia Turística, Thinktur (in Spanish).
Contact	Belén Hidalgo Ferrer: belen.hidalgo@carm.es

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	HODEIAN: Gipuzkoa Smart Destination Data Analytics
Main Stakeholder	Gipuzkoa Provincial Council (Gipuzkoako Foru Aldundia)
Location	Spain
Description	HODEIAN is a tool analysing consumer habits, and how visitors move around the territory of Gipuzkoa. It aims to be an aid in decision-making for public/private managers, tourism resources and any commercial establishment that may be related to tourism. The HODEIAN project is being financed by the Provincial Council of Gipuzkoa. Some of the specific objectives being pursued are: To find out quantitative and qualitative data on day-trippers (visitors who do not stay overnight) who are outside the traditional tourism statistics. To find out how visitors are distributed and move around the territory of Gipuzkoa, where they spend the night and where they carry out their main activity during the day. To find out visitors' purchasing habits by their profile of origin and by the type of expenditure.
Data	Monitoring of expenditure in commercial establishments: Data is obtained on transactions carried out throughout Gipuzkoa with cards from a bank and with Point of Sale terminals from the same entity. Monitoring of visitor movements: Data is obtained from a mobile phone operator on the mobility of visitors throughout Gipuzkoa according to their profile. Counting of people through sensors: Data is obtained on the number of people visiting tourist resources or significant points of interest in Gipuzkoa.
Link	http://www.hodeian.eus/es/index.php
Contact	Jesús Herrero: jesus.herrero@tecnalia.com

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Tourism Intelligence System
Main Stakeholder	SEGITTUR
Location	Spain
Description	The Tourism Intelligence System (Sistema de Inteligencia Turística, SIT, in Spanish) is proposed as an instrument based on the exhaustive analysis of different sources of information, selected according to the needs and idiosyncrasies of the territory. This model of tourism information management, developed using "Business Intelligence" and "Big Data" technology, is capable of automatically generating useful, valid and reliable information, putting it at the service of the actors in the tourism destination, facilitating the dissemination of knowledge and contributing to improving the strategic planning processes of tourism destinations. The System can be adapted to each destination, and it has been developed under proprietary software (Microsoft Power BI), as well as open-source software (Pentaho). Several destinations are currently implementing the system: Las Palmas, Mallorca, Badajoz-Elvas, and Buenos Aires (Argentina). The Tourism Intelligence System is part of the "Smart Destinations" initiative from SEGITTUR. Other planned projects related to this initiative are: SMARTUR, a technological platform for the management of Smart Tourist Destinations; and DATAESTUR, a website that will collect basic data on tourism in Spain and from which it will be possible to access the various sources of tourism statistics, mainly from public and private organisations.
Data	The Tourism Intelligence System is capable of facilitating the analysis of a large volume of heterogeneous information, being able to analyse this information in real time, substantially reducing the economic cost associated with data collection and integrating (alongside traditional data sources), the new data sources based on digital environments: credit cards, social networks, OTAS, sensors, mobiles, groupware etc.
Link	https://www.segittur.es/transformacion-digital/proyectos-transformacion-digital/sistema-de-inteligencia-turistica-2/
Contact	prensa@segittur.es

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Name of the initiative	Tourism Intelligence System: Badajoz-Elvas
Main Stakeholder	Badajoz City Council
Location	Spain
Description	The project of a shared Tourism Intelligence System in the cross-border area of Badajoz-Elvas (Spain-Portugal) was developed by the City Council of Badajoz, the Municipal Chamber of Elvas, and SEGITTUR, and it was co-financed with ERDF funds in the framework of the Spain-Portugal Cross-Border Cooperation Programme (POCTEP). Developed with Pentaho.
Data	Economic data: consumption and expenditure resulting from economic activity obtained from data from BBVA cards and transactions originating in BBVA POS terminals, from any card of any entity and nationality. Online data: Events that take place at the destination. Forecast of hotel demand. Positioning against competitors. Badajoz-Elvas data: Museum tickets according to time period. Tickets to clinics according to time period. People attended in tourist information offices according to period of time. Mobility data: Entrance to car parks according to time period. Visits and movement of visitors according to time period. Visits, movement of visitors and average time according to points of interest. Ranking of the most visited routes according to points of interest.
Link	http://www.sitbadajozelvas.es/
Contact	prensa@segittur.es

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

3 Recommendations and Conclusions

The incorporation of big data sources in the tourism sector opens up a new range of possibilities for destinations to improve the services they offer and the management of their tourism products. This study is a step towards understanding the possibilities of big data in tourism at a Mediterranean level. As mentioned, the introduction of big data is in its early stages in the tourism sector. However, its significant capacity to complement tourism statistics, and its potential to provide real-time information, has been widely recognised. In essence, the optimal approach would be to expand the current state of tourism statistics into a framework that successfully integrates data from traditional tourism surveys, administrative databases, and new sources of big data.

As such, this section aims to provide an overview of the most relevant recommendations on the use of new data sources in the tourism sector. These general suggestions are based on the previously described big data initiatives in tourism at the Mediterranean level, as well as big data projects outside the region, and the growing scientific literature on tourism and big data. Considering the scope of the project, these recommendations have been structured according to the type of big data source, and the feasibility of their implementation.

3.1 User-Generated Content (UGC) Data

UGC data in the context of tourism encompasses mostly online textual data, such as online reviews and opinions from platforms like TripAdvisor, Booking, Expedia, Airbnb, Twitter, Facebook, etc.; and online geo-tagged photos shared on social media. In order to correctly take advantage of the useful knowledge concealed in online data, we suggest the use of the following general approach:

Step 1: Define the problem and the type of data that is required. If the purpose is to measure tourist satisfaction, then online review data for hotels, restaurants, and tourist attractions have proven to be useful in determining tourists' perceptions. But if the purpose is to collect tourism recommendations or conduct sentiment analysis, then it is advisable to use data from social media sites such as Twitter or travel blogs.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Step 2: Collect online textual data from related social media websites, booking platform reviews, and travel blogs. In general, the most popular data source for online reviews is TripAdvisor, while the two main sources of social media data are Twitter and Sina Weibo (China's alternative to Facebook and Twitter). When available, the preferred approach is to use APIs, which standardise the process of directly collecting web data, while also allowing the provider to maintain control over data access. In the context of tourism-related websites, several APIs are freely available, for instance, Twitter API, Facebook Graph API, Flickr API, and Sina Weibo API. In case APIs are not openly available, hidden APIs can sometimes be accessed for scraping data, but most of the time it will be necessary to develop a web scraper or web crawler that extracts the necessary information directly from the website (for data protection recommendations and policies see paragraph d).

Step 3: Pre-process and analyse the data. As far as data pre-processing is concerned, in general the following procedures are considered as standard: data cleaning, tokenization, word stemming, and part-of-speech tagging. Once the data has been cleaned, it is recommended to implement analysis techniques such as:

a) Sentiment analysis can be used to discover tourists' attitudes to services, attractions or destinations by classifying textual data into categories such as positive, negative or neutral, or also particular emotions such as anger, anticipation, disgust, fear, joy, sadness, surprise, or trust.
b) Basic statistical procedures can also be applied to textual data to understand the most frequent terms used by the tourists, the correlation between terms, term co-occurrences, etc.
c) Other more advanced techniques can also be adopted at a later stage, for instance, latent Dirichlet allocation (LDA), a model for extracting meaning from textual data, can be used to identify and label the topics that determine tourist satisfaction from TripAdvisor review data; clustering and categorization can be applied to group reviews based on topics or geographical location; text summarisation has proven to be useful to identify the key aspects in online reviews; and dependency analysis has been successful in establishing the relationship between online reviews and the performance of tourism establishments.

In broad terms, social media data represents one of the most feasible sources of big data for short term implementation, due to its public availability and the growing

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

number of users that continuously generate content. Therefore, it is advisable to consider the following general recommendations:

a) To ensure relevant results limit the dataset to social media content that is time-stamped and geo-located. In this way it is also possible to guarantee a more accurate linkage of social media data with other data sources.
b) Sentiment analysis results can also be clustered in terms of points of interest (POI), namely tourist attractions and accommodations. As such, it is possible to obtain better insights into the reasons that attract tourists to certain specific places.
c) In relation to online image data, Flickr is the most commonly used source for information extraction. In general, it is recommended to consider the following metadata for obtaining meaningful results: photo ID, user ID, date and time, geotag information, title, image description, and tags. Online photo data has been used successfully to provide insights into tourist behaviour, tourist spots, length travel routes, and to present valuable travelling recommendations.
d) Even though social media data is comprised of publicly available content that can be collected from the internet, it is still necessary to guarantee the privacy of individual users and prevent any potential violations. For this reason, it is recommended to follow the guidelines established in *ESSnet's Web Scraping Policy*, where it is stated that data should be collected in a reasonable and ethical way that avoids putting pressure on website owners, while also complying with the EU's General Data Protection Regulation to avoid scraping any personal or copyrighted data.

Examples:

HERIT-DATA: Innovative Solutions to Better Manage Tourism Flows Impact on Cultural and Natural Heritage Sites through Technologies and Big Data, Tuscany Region.

Big Data & Tourisme, De Vinci Research Center.

Monitoring and Listening to the Web and Social Media through Social Analytics and Social Listening Services, Italian National Tourism Board (ENIT).

SmartData, Turismo de Andalucía.

Tourism Statistics and the Use of Social Media Data, Statistics Netherlands (CBS).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

3.2 Operations Data

3.2.1 Web Search Data

The most common web search data sources used in tourism are Google Trends (online search frequency for a particular term in Google) and Baidu index (online search frequency for China's largest search engine). Google Trends data is recommended to be used for most tourist destinations and markets due to its extensive geographical scope. Google Trends data is publicly available, easily accessible, and simple to analyse, making it a source of big data that is feasible to implement in the short term regardless of the level of technological innovation of the destination. However, for the estimation of tourist demand from China the use of the Baidu index is strongly advised.

Regarding the application of techniques to extract information from web search data, the following general approach is recommended:

a) Select relevant keywords with respect to the destination or tourist markets under consideration. The quality of the results depends critically on the selection of the most appropriate keywords. This process can be carried out through directly selecting keywords based on empirical experience, limiting keywords to specific tourist attractions or POI, or selecting keywords based on their predictive capabilities.
b) Raw or indexed web search data can be used to forecast tourism demand, hotel demand, daily tourist volume, tourist flows and inflows, with simple regression techniques.

Examples:

SmartData, Turismo de Andalucía.

3.2.2 Online Booking Data

Online booking data can be helpful in assessing the purchasing behaviour of tourists and the economic impact of different accommodation structures. Still, most of the efforts being made in this area are related to the extraction of data from booking

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

platforms, and their subsequent integration into official statistics, in order to quantify the effect of the sharing economy in the tourism sector of a particular destination.

As already mentioned, the preferred approach to extract online booking data is to use APIs, but in case they are not available, it is necessary to develop a web scraper or web crawler that extracts the necessary information directly from the website. Unfortunately, this is the case for most booking platforms such as TripAdvisor, Booking, and Airbnb.

In particular, the following general recommendations can be considered:

a) Active listings for lodgings can be accessed by scrapping booking platforms at several points in time.
b) The number of bookings per accommodation and time period can be estimated based on the number of reviews. Since not all guests leave online reviews on booking platforms, this can be solved by dividing the number of total reviews by the review rate for each website. For example, Airbnb has stated that its review rate is 72%, and Statistics Netherlands has suggested that Booking.com's review rate is around 24%.
c) Data regarding the number of guests can be roughly estimated considering the capacity of the accommodation and its occupancy rate. Or in the case of Booking.com, for example, by extracting the information concerning the type of guests: solo guest, couple, family with children, and group of friends.
d) Avoid using simple address information for linking data, since it is prone to errors and inconsistencies, and it is often not listed in many booking platforms. Instead, a distance-based approach can be applied, using latitude and longitude information in combination with data on tourist accommodation establishments from different sources (accommodation surveys, for instance), as tested during the ESSnet Big Data II: WPJ - Innovative Tourism Statistics project. Additionally, fuzzy linkage based on accommodation names can be sufficient if dealing with rural destinations, but not in the case of other types of destinations. It is also recommended to perform data linkage based on additional variables such as business register data, accommodation capacity, normalized number of reviews, and the text description of the accommodation.
e) It is essential to implement deduplication techniques to eliminate accommodations that are present on more than one of the booking platforms.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

f) Once again, it is recommended to follow the guidelines established in *ESSnet's Web Scraping Policy* and comply with the EU's General Data Protection Regulation to avoid scraping any personal or copyrighted data.

A variety of projects have been dedicated to extracting and openly distributing data from various online booking sites, representing a simpler, but less comprehensive alternative to webscraping, for example: Inside Airbnb, DataHippo, and Airbnb vs. Berlin.

In addition, it is important to mention the recent agreement established by the European Commission with Airbnb, Booking, Expedia, and TripAdvisor for the upcoming publication of data through Eurostat. As a result of this agreement, aggregated data containing the number of bookings and the number of guests per municipality for the four platforms is expected to be openly published.

Other interesting initiatives currently under development that try to breach the gap concerning short-term rentals data are TripAdvisor Navigator, a proprietary data-intelligence solution for improving destination marketing; and Airbnb's City Portal, a solution developed for government entities that aims to provide data related to Airbnb usage by city.

Examples:

ESSnet Big Data II: WPJ - Innovative Tourism Statistics, Eurostat.

Measurement of the Number of Tourist Dwellings in Spain and their Capacity, National Statistics Institute (INE), Spain.

Estimation of the Number of Guests and Overnight Stays in Platform-Related Accommodations, Statistics Netherlands (CBS).

Holiday Rentals Statistics, Statistics Finland.

3.2.3 Card Transaction Data

When tourists make payments within the destination, transaction data are recorded. These data contain valuable information for determining tourists' purchasing habits and for developing improved products and services. In general, these transactions can include credit card purchases made through Point of Sale (POS) as well as cash withdrawals from ATMs. Accessing card transaction data involves establishing direct

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

cooperation with data providers, in this case, the ESSnet quality guidelines for the acquisition and usage of big data in the case of direct cooperation should be followed, specifically:

a) The following areas should be thoroughly examined: the population coverage, the units of measurement, variables, timeliness and frequency, and information on the organisation.
b) Data file requirements must be addressed in a professional manner, in particular the means for data access, and in the case of pre-processed data access, the clarity of the technical processes applied to the data, the transmission time, and the metadata should be clearly specified.
c) Long-term access must be guaranteed in order to avoid coverage and compatibility errors.
d) Governance issues, including a change in management and dispute resolution mechanisms need to be addressed.

Furthermore, the following general recommendations for the use of card transaction data can be established:

a) Results should be interpreted consistently. Transaction data provide information on expenditure per card, i.e. expenditure per family unit and not necessarily expenditure per tourist.
b) The key information that should be extracted is related to average expenditure per card accumulated throughout the stay, average daily expenditure, average expenditure per card accumulated throughout the stay per tourist establishment, and average daily expenditure per tourist establishment.
c) Card transaction data can be used to provide key insights on the possible types of tourism within a destination, the services required by different types of tourists, the distribution of tourism expenditure across the territory, the impact of tourism on economic development, etc.
d) Privacy concerns should be carefully addressed. Data providers have addressed potential privacy concerns regarding the use of card transaction information by only providing aggregated and anonymised data.

Among the big data initiatives presented in this report that use transaction data, most use data obtained from direct cooperation with Mastercard. In fact, Mastercard is carrying out a number of tourism-related initiatives such as the publication of the Global

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Destination Cities Index, and Tourism Insights, a proprietary solution for government agencies. In the case of Spain, the transaction data used to generate information on tourism expenditure is usually the product of direct cooperation with the BBVA bank.

Examples:

HODEIAN: Gipuzkoa Smart Destination Data Analytics, Gipuzkoa Provincial Council (Gipuzkoako Foru Aldundia), Spain.

Tourism Intelligence System: Badajoz-Elvas, Badajoz City Council, Spain.

Distribution of the Expenditure made by Foreign Visitors on Visits to Spain, National Statistics Institute (INE), Spain.

Consumption of Foreign Tourists – Experimental, Statistics Iceland.

Big Data y Turismo, SECTUR, Mexico.

Monthly Regional Tourism Estimates (MRTEs), Ministry of Business, Innovation & Employment of New Zealand.

Tourist Spending Insights, Center for International Development at Harvard University, USA.

3.3 Device Data

GPS, MNO, Bluetooth, RFID, and Wi-Fi data have been used to monitor the mobility of tourists, and these data sources have the potential to produce quality data for the tourism industry even for less visited destinations or crowded locations. Among the big data initiatives presented in this report, there is a marked trend towards exploratory projects related to the use of MNO data.

Similar to the case of card transaction data, using MNO data involves direct cooperation with the data provider, thus the same quality guidelines should be followed, but in combination with specific recommendations established by the ESSnet Big Data II project:

a) Agree with MNOs on the different roles played by MNOs and the organisation.
b) Agree with MNOs on the raw data to be used in the statistical process, paying special consideration to the technological and business feasibility to use these data.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

c) Agree with MNOs on the statistical processing of raw data to generate intermediate data for the further analysis.
d) Document both which raw data should be used and the data pre-processing procedures. This documentation should consider public transparency, privacy, confidentiality, and intellectual property rights (Eurostat, 2020d).

Additionally, further recommendations can be drawn for the use of MNO data, such as:

a) The number of active foreign cell phones can be used as a basis for extracting information on the origin of tourists, their territorial preferences, their average stay, and their movements within the tourist destination or between different cities.
b) Because of low geographical precision, MNO data are not appropriate for analytical purposes where the position of tourists has to be precise. For instance, MNO data has proven to be successful for measuring the length of stay, but it cannot determine the exact accommodation establishment where the tourist has stayed.
c) Since most mobile positioning data initiatives for tourism statistics are managed separately by different stakeholders in multiple countries, one specific solution being tested in Nordic countries is the development of a regional knowledge transfer and collaboration network for MNO data initiatives (Nordic Innovation & Swedish Agency for Regional and Economic Growth, 2020).

Examples:

Tourism Information Portal (Portal de Informação Turística), NOS, Portugal.

IoT and Big Data in Action: Sagrada Familia, City Council of Barcelona (Ajuntament de Barcelona), Spain.

Measurement of National and Inbound Tourism from the Position of Cell Phones, National Statistics Institute (INE), Spain.

Tourism Intelligence System: Buenos Aires, Tourism Entity of the City of Buenos Aires, Argentina.

Identifying Population Movements Using Anonymised Telephone Data, Statistics Netherlands (CBS).

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

3.4 General Recommendations

Incorporate the use of different data sources. Higher quality tourism information is possible by examining data from a variety of different sources such as big data, administrative databases, and survey data.

Stay constantly up to date with the latest developments in the fields of big data and data science. Similarly, it is necessary to keep the tools used for web scraping, text processing, data mining, and machine learning continuously updated, because most big data sources are not particularly consistent and tend to suffer changes in their structure.

Hire the correct professionals for the proper implementation of big data. At least four separate roles are usually needed to carry out big data projects: a domain expert, a researcher, a computer scientist, and a system administrator. Therefore, a big data project should not be handled by only one person.

Ensure multi-disciplinarity. Big data projects are inherently multi-disciplinary, even more so when it comes to the tourism sector. Depending on the magnitude of the project it is important to involve tourism experts, local stakeholders, methodologists, data scientists, statisticians, economists, etc.

Guarantee the individual privacy of tourists and prevent re-identification. When dealing with highly sensitive data such as mobile positioning, social media, or card transactions, it is critical to ensure the privacy of users through aggregated and anonymized data. As previously mentioned, any big data project should, as a minimum, comply with the guidelines established by ESSnet's Web Scraping Policy and the EU's General Data Protection Regulation.

Avoid high initial computational costs by testing cloud solutions and open source tools. As specified in Table 3, there is a wide range of tools that can be tested before implementing an in-house big data solution. Additionally, the European Commission's Big Data Test Infrastructure (BDTI) initiative is particularly relevant, since it provides EU public administrations with a free cloud-based analytics test environment to experiment with big data technologies.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

3.5 Policy Recommendations

Designate a specific national authority within the tourism sector to be responsible for collecting and standardising data from statistical offices and other public tourism-related organisations, with the aim of publishing all official tourism data in Open Data format and thus guaranteeing ease of access.

This designated authority will then be responsible for making framework agreements with data providers such as MNOs and booking platforms, to ensure nation-wide access to tourism data derived from new sources. Please refer to section 1.3.2.1 for specific guidelines related to the use of big data through direct cooperation with data providers.

This designated authority should be the one responsible for the acquisition of new big data sources related to tourism, and should therefore develop a knowledge transfer and collaboration network or platform where all the different tourism stakeholders can easily access the data, either paid or free of charge.

Example:

Nordic Development Project on Mobile Positioning Data for Tourism Statistics - Nordic Innovation, Denmark.

Project co-financed by the European Regional Development Fund

Glossary

AI – Artificial Intelligence

API – Application Programming Interface

LBSN – Location-Based Social Networks

MNO – Mobile Network Operator

NLP – Natural Language Processing

NSI – National Statistical Institute

RFID – Radio-frequency identification

WPJ – Work Package J (Innovative Tourism Statistics)

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Bibliography

AAPOR. (2015). *AAPOR Report on Big Data*. American Association for Public Opinion Research. https://www.aapor.org/getattachment/Education-Resources/Reports/Big DataTaskForceReport_FINAL_2_12_15_b.pdf.aspx

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, *29*(3), 469–486. https://doi.org/10.1016/j.tourman.2007.05.014

Aureli, S., Medei, R., Supino, E., & Travaglini, C. (2013). Online Review Contents and Their Impact on Three and Four-Star Hotel Reservations: Some Evidence in Italy. In Z. Xiang & I. Tussyadiah (Eds.), *Information and Communication Technologies in Tourism 2014* (pp. 381–393). Springer International Publishing. https://doi.org/10.1007/978-3-319-03973-2_28

Barcaroli, G. (2015). *Use of Big Data in official statistics*. ISI World Statistics Congress, Rio de Janeiro. https://www.researchgate.net/publication/281244223_Use_of_Big_Data_in_official_statistics

Batista e Silva, F., Marín Herrera, M. A., Rosina, K., Ribeiro Barranco, R., Freire, S., & Schiavina, M. (2018). Analysing spatiotemporal patterns of tourism in Europe at high-resolution with conventional and big data sources. *Tourism Management*, 15. https://doi.org/10.1016/j.tourman.2018.02.020

Braaksma, B., & Zeelenberg, K. (2020). *Big data in official statistics* (p. 23) [Discussion paper]. Statistics Netherlands (CBS). https://www.cbs.nl/-/media/_pdf/2020/04/dp-big-data-in-official-statistics.pdf

Braaksma, B., Zeelenberg, K., & De Broe, S. (2020). Big Data in Official Statistics: A Perspective from Statistics Netherlands. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big Data Meets Survey Science* (1st ed., pp. 303–338). Wiley. https://doi.org/10.1002/9781118976357.ch10

Centobelli, P., & Ndou, V. (2019). Managing customer knowledge through the use of big data analytics in tourism research. *Current Issues in Tourism*, *22*(15), 1862–1882. https://doi.org/10.1080/13683500.2018.1564739

Chessa, A., Verburg, J., & Willenborg, L. (2017). *A comparison of price index methods for scanner data*. 31.

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Cortina García, F., Izquierdo Valverde, M., Prado Mascuñano, J., & Velasco Gimeno, M. (2016). Quality implications of the use of big data in tourism statistics: Three exploratory examples. *European Conference on Quality in Official Statistics (Q2016)*, 11. https://www.ine.es/q2016/docs/q2016Final00015.pdf

Daas, P. J. H., Puts, M. J., Buelens, B., & Hurk, P. A. M. van den. (2015). Big Data as a Source for Official Statistics. *Journal of Official Statistics*, *31*(2), 249–262. https://doi.org/10.1515/jos-2015-0016

Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. (2018). Open Innovation and Social Big Data for Sustainability: Evidence from the Tourism Industry. *Sustainability*, *10*(9), 3215. https://doi.org/10.3390/su10093215

Demunter, C. (2017). *Tourism statistics: Early adopters of big data? 2017 edition*. Publications Office of the European Union. http://dx.publications.europa.eu/10.2785/762729

Erl, T., Khattak, W., & Buhler, P. (2016). *Big data fundamentals: Concepts, drivers & techniques*. Prentice Hall.

European Commission. (2019). *CEF Big Data Test Infrastructure Service Offering Description*. CEF Digital. https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/Test+Infrastructure?preview=/82773753/132350224/Test%20Infrastructure%20(ServiceOfferingDescription)%20(v1.00)%20-%20final%20version.pdf

European Commission. (2020). *ESSnet Big Data1* [Wiki]. ESSnet Big Data. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data1

Eurostat. (2012). *Report on modified Work Package 5: The Architecture and applications to handle the XML files received directly from the management systems* (ESSnet on Automated Data Collection and Reporting in Accommodation Statistics). https://ec.europa.eu/eurostat/cros/content/final-report-work-package-5_en

Eurostat. (2019a). *Deliverable J1: Methods for webscraping data processing and analyses* (WPJ: Innovative Tourism Statistics). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/4/4f/WPJ_Deliverable_J1_ESSnet_Methods_for_webscraping_data_processing_and_analyses_2019_07_23.pdf

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Eurostat. (2019b). *Milestone JM1: Report on the WPJ meeting in Rzeszów (PL) on 8-10 July 2019* (WPJ: Innovative Tourism Statistics). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/23/WPJ_Milestone_JM1_Meeting_2019_07_08-10_Rzesz%C3%B3w_Minutes.pdf

Eurostat. (2019c). *Milestone LM3: Final report*. ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e9/WPL_Milestone_LM3_Final_Report_2019_10_31.pdf

Eurostat. (2020a). *Deliverable J2: Interim technical report showing the preliminary results and a general description of the methods used* (WPJ: Innovative Tourism Statistics). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/6/60/WPJ_Deliverable_J2_-_Interim_technical_report_showing_the_preliminary_results_and_a_general_description_of_the_methods_used_2020_01_07.pdf

Eurostat. (2020b). *Deliverable J3: Methodological framework report* (WPJ: Innovative Tourism Statistics). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/3/3d/WPJ_Deliverable_J3_Methodological_framework_report_2020_03_13.pdf

Eurostat. (2020c). *Deliverable J4: Technical Report* (WPJ: Innovative Tourism Statistics). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/2/21/WPJ_Deliverable_J4_Technical_report_2020_06_16_final.pdf

Eurostat. (2020d). *Deliverable K3: Revised Version of the Quality Guidelines for the Acquisition and Usage of Big Data* (WPK: Methodology and Quality). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/f/f8/WP3_Deliverable_K3_Revised_Version_of_the_Quality_Guidelines_for_the_Acquisition_and_Usage_of_Big_Data_Final_version.pdf

Eurostat. (2020e). *Deliverable K5: First draft of the methodological report* (WPK: Methodology and Quality). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/a/a0/WPK_Deliverable_K5_First_draft_of_the_methodological_report_2020_06_17_Finalv3.pdf

Eurostat. (2020f). *Deliverable K7: Typification matrix for big data projects* (WPK: Methodology and Quality). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/e/e0/WPK_Deliverable_K7_Typification_matrix_for_big_data_projects_2020_04_30.pdf

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Eurostat. (2020g). *Milestone JM2: Report on the virtual WPJ meeting on 7 and 20 July 2020* (WPJ: Innovative Tourism Statistics). ESSnet Big Data II. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/images/7/7d/WPJ_Milestone_JM2_Meeting_2020_07_9_%26_20_Virtual_11_Minutes.pdf

Feldman, R., & Sanger, J. (2007). *The text mining handbook advanced approaches in analyzing unstructured data*. Cambridge University Press. https://wtlab.um.ac.ir/images/e-library/text_mining/The%20Text%20Mining%20HandBook.pdf

Ferreira Dinis, M. G., Martins da Costa, C. M., & da Rocha Pacheco, O. M. (2019). Composite Indicators for Measuring the Online Search Interest by a Tourist Destination. In M. Sigala, R. Rahimi, & M. Thelwall (Eds.), *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-6339-9

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing & Management*, *3*(4), 198–209. https://doi.org/10.1016/j.jdmm.2014.08.002

Ghotkar, M., & Rokde, P. (2016). Big Data: How it is Generated and its Importance. *National Conference on Recent Trends in Computer Science and Information Technology (NCRTCSIT-2016)*, 1–5. http://www.iosrjournals.org/iosr-jce/papers/conf.15013/Volume%202/1.%2001-05.pdf

González Gómez, S., & Rubio Gil, Á. (2020). Análisis bibliométrico de big data en el entorno de la generación del conocimiento del turismo. *Revista Internacional de Organizaciones*, *24*, 211–239. https://doi.org/10.17345/rio24.211-239

Japec, L., & Lyberg, L. (2020). Big Data Initiatives in Official Statistics. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big Data Meets Survey Science* (1st ed., pp. 273–302). Wiley. https://doi.org/10.1002/9781118976357.ch9

Joseph, G., & Varghese, V. (2019). Analyzing Airbnb Customer Experience Feedback Using Text Mining. In M. Sigala, R. Rahimi, & M. Thelwall (Eds.), *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-6339-9

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management*, *68*, 301–323. https://doi.org/10.1016/j.tourman.2018.03.009

Li, X., & Law, R. (2020). Network analysis of big data research in tourism. *Tourism Management Perspectives*, *33*, 100608. https://doi.org/10.1016/j.tmp.2019.100608

Miah, S. J., Vu, H. Q., Gammack, J., & McGrath, M. (2016). A big data analytics method for tourist behaviour analysis. *Information and Management*, 43. http://dx.doi.org/doi:10.1016/j.im.2016.11.011

Nordic Innovation, & Swedish Agency for Regional and Economic Growth. (2020). *Nordic development project on Mobile Positioning Data for Tourism Statistics*. https://www.norden.org/en/publication/nordic-development-project-mobile-positioning-data-tourism-statistics

Pérez Guilarte, Y., & Barreiro Quintáns, D. (2019). Using Big Data to Measure Tourist Sustainability: Myth or Reality? *Sustainability*, *11*(20), 5641. https://doi.org/10.3390/su11205641

Reif, J., & Schmücker, D. (2020). Exploring new ways of visitor tracking using big data sources: Opportunities and limits of passive mobile data for tourism. *Journal of Destination Marketing & Management*, *18*, 100481. https://doi.org/10.1016/j.jdmm.2020.100481

Ruhanen, L., Moyle, C., & Moyle, B. (2019). New directions in sustainable tourism research. *Tourism Review*, *74*(2), 138–149. https://doi.org/10.1108/TR-12-2017-0196

Ruhanen, L., Weiler, B., Moyle, B. D., & McLennan, C. J. (2015). Trends and patterns in sustainable tourism research: A 25-year bibliometric analysis. *Journal of Sustainable Tourism*, *23*(4), 517–535. https://doi.org/10.1080/09669582.2014.978790

Samara, D., Magnisalis, I., & Peristeras, V. (2020). Artificial intelligence and big data in tourism: A systematic literature review. *Journal of Hospitality and Tourism Technology*, *11*(2), 343–367. https://doi.org/10.1108/JHTT-12-2018-0118

Syed, A. R., Gillela, K., & Venugopal, D. C. (2013). The Future Revolution on Big Data. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(6), 6.

Project co-financed by the European Regional Development Fund

Tam, S.-M., & Van Halderen, G. (2020). The five V's, seven virtues and ten rules of big data engagement for official statistics. *Statistical Journal of the IAOS*, *36*(2), 423–433. https://doi.org/10.3233/SJI-190595

Valcke, S. (2019). Customer Data and Crisis Monitoring in Flanders and Brussels. In M. Sigala, R. Rahimi, & M. Thelwall (Eds.), *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-6339-9

Vassakis, K., Petrakis, E., Kopanakis, I., Makridis, J., & Mastorakis, G. (2019). Location-Based Social Network Data for Tourism Destinations. In M. Sigala, R. Rahimi, & M. Thelwall (Eds.), *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-6339-9

Volo, S. (2020). Tourism statistics, indicators and big data: A perspective article. *Tourism Review*, *75*(1), 304–309. http://dx.doi.org/10.1108/TR-06-2019-0262

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51–65. https://doi.org/10.1016/j.tourman.2016.10.001

Xiang, Z., Schwartz, Z., Gerdes, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130. https://doi.org/10.1016/j.ijhm.2014.10.013

Xu, F., Nash, N., & Whitmarsh, L. (2019). Big data or small data? A methodological review of sustainable tourism. *Journal of Sustainable Tourism*, *28*(3), 144–163. https://doi.org/10.1080/09669582.2019.1631318

Yamamoto, M. (2019). Furthering Big Data Utilization in Tourism. In F. P. García Márquez & B. Lev (Eds.), *Data Science and Digital Business* (pp. 157–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-95651-0_9

Interreg Mediterranean BEST MED
Project co-financed by the European Regional Development Fund