

Evaluation of Tourist Traceability Alternatives Based on Ubiquitous Systems

Doctoral Thesis



Juan Francisco Mendoza Moreno

Advisor: PhD. Gustavo Adolfo Ramírez González

Universidad del Cauca
Faculty of Electronics and Telecommunications Engineering
PhD in Telematics Engineering
Department of Telematics
Research Line: Advanced Telecommunication Services

Popayán, January 2022

I want to dedicate this thesis to my loving wife Luz and my dear daughter Juana. All the glory to God.

Acknowledgments

I want to express my recognition of thanks to my supervisor, Ph.D. Gustavo Ramírez González for his support in managing this research project. I thank Knowledge Reusing research group at the Carlos III University of Madrid and my research internship tutor, Ph.D. Anabel Fraga Vázquez for guiding me in this research work. I thank the Ph.D. (c) Luz Santamaría Granados for his contribution and improvement in this project. I thank the peer evaluators of this thesis and the reviewers of the published scientific papers for their valuable comments. Finally, thanks to the Ministry of Science, Technology, and Innovation of Colombia (733-2015), Gobernación de Boyacá, Universidad del Cauca, and Universidad Santo Tomás Seccional Tunja for funding the project. Thanks to Fray Eduardo González Gil for being the promoter of my doctoral training. This work could not be possible without any of the people and institutions aforementioned. Therefore my heartfelt thanks for their cooperation.

Abstract

Tourist traceability is the analysis of the set of actions, procedures, and technical measures that identify and record the space-time relationship of the touring through the tourist value chain. Tourist Traceability System (TTS) has implications for infrastructure, transport, products, marketing, and management. A TTS benefits Destination Management Organizations (DMO), which require data for decision-making. The challenge of this research was to analyze and obtain data from different ubiquitous computing sources. Then, data were processed through a TTS model based on Big Data analytics. Thus, the model provides information for decision-making in an agile and timely manner. The development of this model was based on the state-of-the-art analysis to define a conceptual framework of TTS. The framework allowed the analysis of alternatives for tourist traceability. Finally, a TTS model was proposed with these components: a knowledge base called OntoTouTra, location intelligence, decision-making, and visualization. The model was tested and validated through a use case with the generation of Key Performance Indicators (KPI) required for managing DMO. The results of these tests evidenced the robustness and effectiveness of the model in a TTS domain.

Background

Decision-making by the Tourist Destination Managers (DMO) is a critical process for the sustainability of the tourism ecosystem. For this process to succeed, we need to have sufficient and well-updated information. It is difficult for some traditional and emerging tourism systems to obtain the data since many of the sources are manual or dependent on third parties. On the other hand, the amount of data for more developed tourist destinations is enormous for processing.

If third-party data processing systems are not connected online with DMO systems, we will miss valuable opportunities in managing our destination.

In most developing countries, decision-making for tourism systems depends mainly on statistical data from airports, terminals, tolls, hotel occupancy, restaurants, among others.

Most of the information is processed manually or semi-automatically and is available over long periods, annually, semi-annually, or in the best of cases, monthly. Nevertheless, the nature of tourism ecosystems requires that this data be available in real-time or near real-time. Besides, we expect other essential data, such as the concurrence of the destination's Points of Interest (PoI) and the tourist's behavior while traveling the destination. We considered the history of the tourist movement as tourist traceability.

At the moment, tourism traceability systems have not been addressed in depth. Most studies deal with tracking or tracking the tourist in some destination places, such as hotels, theme parks, and museums. Traceability allows us to reconstruct the tourist's history, route, and interaction to identify the origin of their visit, the history of the tourist's interaction with the destination and its actors, and the location during and after the visit.

It is challenging to apply a tourist traceability system for the management of the destination because real-time data is required, coming from diverse and varied data sources, to consolidate a knowledge base accessible not only to the actors but also understandable among the data processing machine and of course, the processing of personal data and other sensitive data.

Thus, in this research, we have detected the lack of a tourism traceability model that takes advantage of ubiquitous data sources and consolidates a knowledge base for proper decision-making by DMO.

Aims

This project aims to propose an alternative of tourist traceability based on ubiquitous platforms for designing the experiences of a tourist destination. The following specific objectives were required:

- Structure a dataset according to context-aware ubiquitous systems and tourist opinion data.
- Design the tourist traceability model using Big Data predictive analysis techniques.
- Validate the Big Data predictive analytical model for the generation of the tourist experiences design of the destination.

Method

The literature review analyzed the research trends related to tourism traceability. We delved into the location, the techniques to obtain it, and associated technologies. Then,

the analysis of knowledge bases associated with tourism management, mainly ontologies. Finally, the study of ubiquitous data sources' influence on tourism systems, especially social media.

Later we developed a web scraping system to obtain data from tourism-oriented social networks, especially online tourism agencies (OTA).

We also developed an application that we installed on the participants' smartphones of an experimental group of researchers. This application generated location and interaction data with the sensors in an experimental tourist destination. The sensors used were beacons.

To process this data, we used Big Data analytical techniques, and as a result, we generated the tourism traceability model based on ubiquitous data.

Thanks to this model, we generate a knowledge base. It is an ontology that we call *OntoTouTra*, which resides in an end-point that we create. Finally, we created a module to visualize tourism traceability data and a generator module for a portfolio of tourism experiences for the region.

The model was validated with data from the experiment and the case study, with actual tourism data.

Results

The validation results found that the ubiquitous data source that contributed the most to the ontology is social media. The words belong to the specific domain of tourism and are not exempt from the influence of social media. It is an informal, abbreviated language that follows the communicational parameter through conventional digital media. For this reason, for its treatment, it was necessary to use Natural Language Processing (NLP) techniques, as demonstrated by the experiment of determining the polarity of tourist reviews.

The sensors installed in the destination and the tourist's mobile devices provide us with the remaining percentage of the data, especially those related to the location and time series, necessary to determine the traceability of the tourist.

Conclusions

We presented an alternative model of tourism traceability systems based on ubiquitous data sources. The input of this model has three possible sources: social media, sensor data, and apps. The data are acquired and processed with Big Data analytical techniques and make up a knowledge base: The *OntoTouTra* ontology. From this ontology, we generated a

module of traceability visualizations and another module to create the portfolio of tourist experiences of the destination. The applications of this model are comprehensive, but it is mainly aimed at generating information that supports decision-making by DMO.

Keywords:

Tourist traceability, ontology, OntoTouTra, ubiquitous systems, Big Data analytic, social media, sensors, beacons, location.

Resumen

La trazabilidad turística es el análisis del conjunto de acciones, procedimientos y medidas técnicas que identifican y registran la relación espacio-tiempo del turismo a través de la cadena de valor turística. El Sistema de Trazabilidad Turística (TTS) tiene implicaciones para la infraestructura, el transporte, los productos, la comercialización y la gestión. Un TTS beneficia a las Organizaciones de Gestión de Destinos (DMO), que requieren datos para la toma de decisiones. El desafío de esta investigación fue analizar y obtener datos de diferentes fuentes informáticas ubicuas. Luego, los datos se procesaron a través de un modelo TTS basado en analíticas de Big Data. Así, el modelo brinda información para la toma de decisiones de manera ágil y oportuna. El desarrollo de este modelo se basó en el análisis del estado del arte para definir un marco conceptual de TTS. El marco permitió el análisis de alternativas para la trazabilidad turística. Finalmente, se propuso un modelo TTS con estos componentes: una base de conocimiento denominada OntoTouTra, inteligencia de ubicación, toma de decisiones y visualización. El modelo fue probado y validado a través de un caso de uso con la generación de indicadores clave de rendimiento (KPI) necesarios para administrar DMO. Los resultados de estas pruebas evidenciaron la robustez y efectividad del modelo en un dominio TTS.

Antecedentes

La toma de decisiones por parte de los gestores de destinos turísticos (DMO) es un proceso crítico para la sostenibilidad del ecosistema turístico. Para que este proceso sea exitoso, necesitamos tener información suficiente y bien actualizada. Para algunos sistemas turísticos tradicionales y emergentes, es difícil obtener los datos, ya que muchas de las fuentes son manuales o dependen de terceros. Por otro lado, para los destinos turísticos más desarrollados, la cantidad de datos para su procesamiento es enorme. Si los sistemas de procesamiento de datos de terceros no están conectados en línea con los sistemas DMO, perderemos valiosas oportunidades en la gestión de nuestro destino. En la mayoría de los países en desarrollo, la toma de decisiones para los sistemas turísticos depende

principalmente de datos estadísticos de aeropuertos, terminales, peajes, ocupación hotelera, restaurantes, entre otros. La mayor parte de la información se procesa de forma manual o semiautomática y está disponible durante largos períodos, anualmente, semestralmente o, en el mejor de los casos, mensualmente. Pero la naturaleza de los ecosistemas turísticos requiere que estos datos estén disponibles en tiempo real o casi en tiempo real. Además, esperamos otros datos que son muy importantes, como la concurrencia de los Puntos de Interés (PoI) del destino y el comportamiento del turista en su recorrido por el destino. Vamos a considerar la historia del movimiento turístico como trazabilidad turística. Por el momento, los sistemas de trazabilidad turística no se han abordado en profundidad. La mayoría de los trabajos tratan sobre el rastreo o seguimiento del turista, en algunos lugares del destino, como hoteles, parques temáticos, museos, entre otros. La trazabilidad nos permite reconstruir la historia, la ruta y la interacción del turista, de tal forma que podamos identificar el origen de su visita, la historia de la interacción del turista con el destino y sus actores; y la ubicación durante y después de su visita. Es muy difícil aplicar un sistema de trazabilidad turística para la gestión del destino porque se requieren datos en tiempo real, provenientes de diversas y variadas fuentes de datos, para consolidar una base de conocimiento accesible no solo a los actores sino también comprensible entre los procesadores de datos. máquina y, por supuesto, el procesamiento de datos personales y otros datos sensibles. Así, en esta investigación hemos detectado la falta de un modelo de trazabilidad turística que aproveche las fuentes de datos ubicuas y consolide una base de conocimiento para la correcta toma de decisiones por parte de las OGD.

Objetivos

Este proyecto tiene como objetivo es proponer una alternativa de trazabilidad turística basada en plataformas ubicuas, para el diseño de experiencias de un destino turístico. Los objetivos específicos son:

- Estructurar un conjunto de datos de acuerdo con sistemas ubicuos sensibles al contexto y datos de opinión turística.
- Diseñar el modelo de trazabilidad turística utilizando técnicas de análisis predictivo Big Data.
- Validar el modelo analítico predictivo Big Data para la generación del diseño de experiencias turísticas del destino.

Métodos

En la revisión de la literatura, analizamos las tendencias de investigación relacionadas con la trazabilidad turística. Luego, profundizamos en temas relacionados con la ubicación, las técnicas para obtenerla y tecnologías asociadas. Luego, el análisis de las bases de conocimiento asociadas a la gestión turística, principalmente ontologías. Finalmente, el estudio de la influencia de las fuentes de datos ubicuas para los sistemas turísticos, especialmente las redes sociales. Posteriormente desarrollamos un sistema de web scraping para obtener datos de redes sociales orientadas al turismo, especialmente de agencias de turismo online (OTA). También desarrollamos una aplicación que instalamos en los smartphones de los participantes de un grupo experimental de investigadores. Esta aplicación generó datos de ubicación e interacción con los sensores que localizamos en un destino turístico experimental. Los sensores utilizados fueron balizas. Para el procesamiento de estos datos, utilizamos técnicas analíticas de Big Data, y como resultado, generamos el modelo de trazabilidad turística basado en datos ubicuos. Gracias a este modelo generamos una base de conocimiento, es una ontología que llamamos OntoTouTra, y que reside en un end-point que creamos. Finalmente, creamos un módulo para la visualización de datos de trazabilidad turística y un módulo generador para un portafolio de experiencias turísticas para la región. El modelo fue validado con datos del experimento y con el estudio de caso.

Resultados

En los resultados de la validación, encontramos que la fuente ubicua de datos que aportó más términos a la ontología son las redes sociales, con un 85 %. Las palabras pertenecen al dominio específico del turismo y no están exentas de la influencia de las redes sociales, es decir, es un lenguaje informal, abreviado que sigue el parámetro comunicacional a través de los medios digitales convencionales. Por ello, para su tratamiento, fue necesario utilizar técnicas de Procesamiento del Lenguaje Natural (NLP), como demuestra el experimento de determinación de la polaridad de las reseñas turísticas. Los sensores instalados en el destino y los dispositivos móviles del turista nos proporcionan el porcentaje restante de los datos, especialmente los relacionados con la ubicación y las series temporales, necesarios para determinar la trazabilidad del turista.

Conclusiones

Presentamos un modelo alternativo de sistemas de trazabilidad turística basado en fuentes de datos ubicuas. La entrada de este modelo tiene tres fuentes posibles: redes sociales, datos de sensores y aplicaciones. Los datos se adquieren y procesan con técnicas analíticas Big Data y conforman una base de conocimiento: la ontología OntoTouTra. A partir de esta ontología, generamos un módulo de visualizaciones de trazabilidad y otro módulo para crear el portafolio de experiencias turísticas del destino. Las aplicaciones de este modelo son integrales, pero su principal objetivo es generar información que sea el soporte para la toma de decisiones por parte de las DMO.

Palabras clave:

Trazabilidad turística, ontología, OntoTouTra, sistemas ubicuos, analítica de Big Data, redes sociales, sensores, beacons, ubicación.

Contributions of this research

Papers

The papers published as the results of this research are mentioned below:

1. **Mendoza-Moreno, Juan Francisco**; Santamaria-Granados, Luz; Fraga-Vázquez, Anabel; Ramirez-Gonzalez, Gustavo. OntoTouTra: Tourist Traceability Ontology based on Big Data Analytics. Applied Sciences, 2021, 11, 11061. <https://doi.org/10.3390/app112211061>, 1-39.
 - Paper available at: <https://www.mdpi.com/2076-3417/11/22/11061/htm>
 - Journal indexed in: **JCR Q2, SJR Q2, and Publindex A1.**
 - Citations: **1**
 - Contribution for this study: **Tourist traceability system model (This is the core component), Tourist traceability alternatives, results and conclusions**
2. Santamaria-Granados, Luz; **Mendoza-Moreno, Juan Francisco**; Ramirez-Gonzalez, Gustavo. Tourist Recommender Systems Based on Emotion Recognition—A Scientometric Review. Future Internet 2021, 13, 2. <http://doi.org/10.3390/fi13010002>, 1-37.
 - Paper available at: <https://www.mdpi.com/1999-5903/13/1/2/htm>
 - Journal indexed in: **JCR Q2, SJR Q2, and Publindex A2.**
 - Citations: **7**
 - Contribution for this study: **State of the art, Tourist traceability system model (Ubiquitous data sources).**
3. Santamaria-Granados, Luz; **Mendoza-Moreno, Juan Francisco**; Chantre-Astaiza, Angela; Munoz-Organero, Mario; Ramirez-Gonzalez, Gustavo. "Tourist Experiences

Recommender System based on Emotion Recognition with Wearable Data". *Sensors*, 2021, 21 No. 23, 7854, <https://doi.org/10.3390/s21237854>, 1-28.

- Paper available at: <https://www.mdpi.com/1424-8220/21/23/7854/htm>
- Journal indexed in: **JCR Q1, SJR Q2, and Publindex A1.**
- Contribution for this study: **Tourist traceability system model (API design - Open system)**

4. A. F. Hussein, N. Arunkumar, C. Gomes, A. Alzubaidi, Q. Habash, L. Santamaria-Granados, **J. F. Mendoza-Moreno** and G. Ramirez-Gonzalez, "Focal and Non-Focal Epilepsy Localization: A Review," in *IEEE Access*, vol. 6, pp. 49306-49324, 2018, <https://doi.org/10.1109/ACCESS.2018.2867078>.

- Paper available at: <https://ieeexplore.ieee.org/document/8445554>
- Journal indexed in: **JCR Q1, SJR Q1, and Publindex A1.**
- Citations: **32**
- Contribution for this study: **State of the art, Big Data analytics classification**

Ontology and Datasets

The repository of the ontology and its documentation at <http://tourdata.org/>.

The raw dataset of location, which was collected in the experimental phase of this research, is available in the following repository: <https://github.com/luzsantamariag/terser>.

Source Code

The source code to build the OntoTouTra ontology and obtain its individuals (instances) from an OTA is available in the following public repository, including installation instructions: <https://github.com/jfmendozam/ontotoutra>.

Contents

1	Introduction	1
1.1	Problem statement	1
1.2	Objectives	2
1.2.1	General Objective	2
1.2.2	Specific Objectives	3
2	State of the art	5
2.1	Materials and Methods	5
2.2	Tourist Traceability Top Topics	7
2.3	Location	9
2.4	Global Position System (GPS)	11
2.5	Destination	12
2.6	Tourism technologies	13
2.7	Tourism domain ontologies	14
2.8	Tracking	20
2.9	Data Science	21
2.10	Cluster Mapping	21
3	Tourist traceability conceptual framework	25
3.1	Tourism value chain	25
3.2	Methodology for the construction of the conceptual framework	27
3.3	Conceptual framework	29
3.4	Background and overview	30
3.4.1	Traceability	30
3.4.2	Tourist traceability	30
3.4.3	Tourist Traceability System Tourist Traceability System (TTS)	30
3.4.4	Tourist traceability considerations	31
3.4.5	Aims of the implementation of tourism traceability	31

3.4.6	Benefits of tourist traceability	32
3.4.7	Components of tourist traceability	32
3.4.8	TTS information categories	32
3.4.9	Registry keeping	32
3.5	External tourist traceability	33
3.5.1	Applicable businesses	33
3.5.2	Requirements around the tourist	33
3.6	Internal tourist traceability	33
3.6.1	Tourism service traceability	33
3.6.2	Batch identification	34
3.7	Retrieval of traceability information	34
3.7.1	Timeframes	34
3.8	Product units of tourist traceability	34
3.9	Identification codes and marks on TTS	35
3.10	Management of the TTS information	36
3.11	TTS analysis	36
3.12	TTS validation	37
4	Tourist traceability alternatives	41
4.1	Related work	41
4.2	First Tourist Traceability Alternative	47
4.3	Second Tourist Traceability Alternative	49
5	Tourist traceability system model	51
5.1	Ubiquitous data sources	53
5.1.1	Social network data	54
5.1.2	Sensors	54
5.1.3	Mobile app	55
5.2	Location intelligence	56
5.3	Opinion mining	57
5.4	OntoTouTra	57
5.4.1	Introduction	57
5.4.2	Tourist Traceability System	59
5.4.3	OntoTouTra Analysis	59
5.4.4	Development of the Ontology on the Domain of TTS	60
5.4.5	Specification	61
5.4.6	Conceptualization	61

5.4.7	Formalization and implementation	63
5.4.8	Evaluation	64
5.4.9	Documentation	64
5.4.10	Model for the Development of OntoTouTra	65
5.4.10.1	Definition of the ontology's purpose	66
5.4.10.2	Data sources	66
5.4.10.3	Data collecting	66
5.4.10.4	Tourist location dataset	66
5.4.10.5	Tourist reviews dataset	67
5.4.10.6	Ontology input data files	67
5.4.10.7	Ontology building	68
5.4.10.8	Ontology validation	68
5.4.11	Development and Usage of OntoTouTra in Big Data Environments	69
5.4.12	Big Data Analytics Lifecycle for Building the TTS Ontology	69
5.4.12.1	Business case evaluation	70
5.4.12.2	Data identification	71
5.4.12.3	Data acquisition and filtering	71
5.4.12.4	Data extraction	71
5.4.12.5	Data validation and cleansing	72
5.4.12.6	Data aggregation and representation	72
5.4.12.7	Data analysis	72
5.4.12.8	Data visualization	72
5.4.12.9	Utilization of analysis results	73
5.4.13	Using Big Data	73
5.4.13.1	Components of the Analytics Toolkit	73
5.4.13.2	Variety of Data	74
5.4.13.3	Big Data Semantics	75
5.4.13.4	Classification Using Big Data	77
5.5	Making-decision system	78
6	Results	79
6.1	Evaluation	79
6.1.1	Evaluation of the Ontology	79
6.1.2	Conceptual Validation	82
6.1.3	Ontology Testing	83
6.1.4	Analysis of the Results	85

7	Conclusions and future work	87
8	List of Acronyms	91
	References	92
Appendix A	OntoTouTra Implementation - Supplementary Material	110
A.1	OntoTouTra conceptual evaluation	110
A.2	Test cases	111
A.2.1	Test case 1: What percentage of visitors are satisfied with the provider's services?	111
A.2.2	Test case 2: What percentage of users are satisfied with the provider's internet services?	113
A.2.3	Test case 3: Number of daily visitors	114
A.2.4	Test case 4: Impact on the destination of the offer of accommodation companies used by visitors	116
A.2.5	Test case 5: Impact of visits on the destination	118
A.2.6	Test case 6: Influence of accommodation companies in the destination	122
A.2.7	Test case 7: Arrival of foreign tourists (FTA)	126
A.2.8	Test case 8: Inbound and local tourism	128
A.2.9	Test case 9: Seasonality patterns in the destination	131
A.2.10	Test case 10: Portfolio of tourist experiences used	133
A.3	Big Data Analytics Lifecycle for building TTS ontology	134
A.3.1	Definition of the ontology purpose	134
A.3.2	Data Validation & Cleansing	134
A.3.3	Data Aggregation & Representation	135
A.3.4	Data analysis	136
A.3.5	Screenshots of SPARQL queries	138
A.4	Disclaimer	146
Appendix B	Beacons installation	147
Appendix C	MEB App	151
Appendix D	Location intelligence component	153
Appendix E	OntoTouTra Development	155
E.1	Data Treatment	163

List of Figures

2.1	Loaded documents from Scopus and WoS	6
2.2	Research trend in recent years of the top categories.	8
2.3	Top 15 Log-Scatter Graph	9
2.4	Location cluster - Top Log-Scatter Graph	9
2.5	Destination cluster - Top Log-Scatter Graph	12
2.6	Technologies cluster - Top Log-Scatter Graph	13
2.7	Tracking cluster - Top Log-Scatter Graph	20
2.8	Data Science cluster - Top Log-Scatter Graph	21
2.9	Network Visualization	23
3.1	Tourism value chain (Based on [1])	26
3.2	Actors in the tourism value chain (Based on [2])	26
3.3	Features of the tourist value chain (Based on [3])	27
3.4	Conceptual framework for TTS domain	29
3.5	Example of two scenarios of tourist traceability	35
3.6	TTS making decision	39
4.1	Mapping of co-occurrence networks	42
4.2	Clusters of tourist tracking using GPS technologies	43
4.3	Cluster of analysis of tourist mobility	44
4.4	First Tourist Traceability Alternative	49
4.5	Second Tourist Traceability Alternative	50
5.1	TTS model	52
5.2	Tourist traceability system: use case.	60
5.4	OntoTouTra development model.	65
5.5	Big Data lifecycle [4].	70
5.6	Architecture diagram for the data pipeline.	74

A.1	UNWTO - Country Fact Sheets: Colombia - In bound tourism	114
A.2	UNWTO - Country Fact Sheets: Colombia - Accommodation companies	116
A.3	MinCIT - Colombia - Local Arrivals	120
A.4	MinCIT - Colombia - Local Tourism Industry	123
A.5	MinCIT - Colombia - FTA	126
A.6	UNWTO - Colombia - Tourist Seasonality	131
A.7	SPARQL query in OntoTouTra using Apache Fuseki	139
A.8	SPARQL query in OntoTouTra using Apache Jena	140
A.9	SPARQL query in OntoTouTra using Protégé	141
A.10	SPARQL query in OntoTouTra using OpenLink Virtuoso	142
A.11	REST API in OntoTouTra using Fuseki SOH	142
A.12	REST API in OntoTouTra using OBA OpenAPI	143
A.13	OntoTouTra documentation generated by Protégé	144
A.14	OntoTouTra documentation generate by OBA	145
B.1	Locations of the beacons	147
B.2	Cluster of beacons	148
B.3	Map of beacons	148
B.4	Beacons in the POI	149
B.5	Location data objects code	149
B.6	Beacon app	150
C.1	MEB app	151
C.2	MEB dataset	152
D.1	Haversine distance	153
D.2	Manhattan distance	153
D.3	Location dataset	154
D.4	Location operations	154
E.1	Levels of OntoTouTra (using WebVOWL [5])	155
E.2	Web scraping class.	156
E.3	Listing of Data link to GeoNames for obtaining city coordinates	156
E.4	Results of data link to GeoNames for obtaining city coordinates	157
E.5	Score reviews dataset	157
E.6	An example of transformation rules from the Cities spreadsheet	158
E.7	Python code snippet about OTA web scraping	158
E.8	Visualization: Main tourist destinations in Colombia	159

E.9	Tourist destinations in Colombia from OntoTouTra.	160
E.10	Sentiment analysis techniques to determine the Satisfaction KPI	160
E.11	Satisfaction KPI	161
E.12	Polarity and Subjectivity of the reviews	161
E.13	Review data stream: unstructured.	162
E.14	Rating predictor algorithm.	162
E.15	Performance of the rating prediction model.	163

List of Tables

2.1	ScientoPy preprocessing results of the tourism traceability systems dataset	7
2.2	Summary of tourism papers that used GPS technology	12
2.3	Tourism domain ontologies found in the literature review	18
3.1	Example of TTS metadata	37
3.2	Traceability analysis - One Step Back	37
3.3	RTM Example	38
4.1	Some studies of literature review about tourists tracking	45
4.2	Some studies of tourists tracking and trajectory approach	46
5.1	Data sources of the individuals of the main classes of OntoTouTra.	62
5.2	Glossary of a TTS (sample concepts).	63
5.3	OntoTouTra relationships (owl:topObjectProperty).	64
5.4	OTAs (source: Cloudbeds, 2020).	71
5.5	Components of the analytics toolkit.	74
5.6	OntoTouTra statistics.	75
5.7	Tourist review categories.	78
6.1	Applying the goal-question-metric approach from the FOCA methodology.	80
6.2	KPI list.	83
6.3	Expected results.	84
6.4	OntoTouTra features.	85
6.5	OntoTouTra vs. other tourism ontologies.	86

Chapter 1

Introduction

1.1 Problem statement

Destination Management Organizations (DMO) represent organizations within the tourism ecosystem [6], there are groups of stakeholders (especially decision-makers and tourism stakeholders on a national level, and the inhabitants of the destination [7–11]) who need to know and prepare the environment and infrastructure to offer tourists the maximum degree of experience [9, 12, 13] they have chosen in different ways: recreational, diversionary, experiential, experimental, and existential [14]. However, for this decision-making, the managers need to have information about the tourists context and movement in the destination [15, 16].

Traditional ways of obtaining travelers data [9] such as the number of vehicles passing through tolls, data from nearby airports, travel agencies, or hotel occupancy, are not enough to have detailed data [17–20]. Ubiquitous computing [21–27] offers great possibilities [28–30] to provide tourist traceability data [31, 32], while he travels the destination attractions. Once these data have been processed, the destination administrators can know, in addition to the visitors' flow, tourist amount, preferred attractions, duration on site, alternative routes, access difficulties, conglomeration sites, undetected sites, among others. This traceability information, join with other ubiquitous sources data [33], such as social networks (through opinion mining techniques and ratings that the tourist comments about destiny), constitute the input to create systems that allow to DMO's to make decisions with respect to the destination infrastructure and in turn to design the experiences [34–38] according to the tourist expectations [39–41]. Research shows that ubiquitous computing is being used by tourism stakeholders, for example, the Codetur surveys [42], shows that 82% of users prefer it as a destination resources guide and 74% use it as geo-location devices.

In this way, several scientific and technological challenges arise:

1. How to record, in a transparent way, the passage of the tourist while he travels the destination? It is necessary to take advantage of one or more ubiquitous capabilities [43], not only having a sensory layer at the destination, but techniques such as detecting the tourist's location (for example, from the mobile device to obtain the location through GPS systems, services such as LBS, location calculation using techniques such as triangulation based on reference points of WiFi signals, or obtaining location by content tagging services, geo-tagged)
2. How to keep track of the flow of tourists?
3. How to identify the type of experience that the tourist is consuming in the destination?
4. How to detect the pleasure or dislike of consumption of the tourist experience in the destination?
5. After answering the previous questions, how to process that large volume and variety of data, at a high rate of generation?
6. How to facilitate the design of tourist experiences in the destination, based on this traceability data?

There are several research opportunities in the various domains: with regard to ubiquitous computing, the recording and retrieval of historical data according to the context, the management of profile information and the prediction of context information, in this case, design of tourist experiences [44]. From the point of view of traceability, the personal data processing is very sensitive, since in this case, traceability is not applied to objects, on the contrary it is applied to people in heterogeneous environments [45]. This leads to the great challenge that is the design of experiences based on tourist traceability data obtained from data sources of ubiquitous systems, in order to promote a sustainable tourism system [46], although the complexity of the system makes it difficult [47].

1.2 Objectives

1.2.1 General Objective

Propose an alternative of tourist traceability based on ubiquitous platforms, for the experiences design of a tourist destination.

1.2.2 Specific Objectives

- Structure a dataset according to context-aware ubiquitous systems and tourist opinion data.
- Design the tourist traceability model using Big Data predictive analysis techniques.
- Validate the Big Data predictive analytical model for the generation of the tourist experiences design of the destination.

Chapter 2

State of the art

The process of tracing the tourists while consuming their experience in the destination, knowing their profile, their route, the attractions that motivate them, the infrastructure shortcomings, among other situations, is what we will call tourist traceability. In this traceability system, it is necessary to process a constant flow of data in real-time or near. The results of the tourist traceability are beneficial for the Destination Management Operators since they can make decisions about the preparation of the destination. This study analyzes the techniques, technologies, and applications of tourist traceability and its scientific trends

2.1 Materials and Methods

The methodology used is based on the application of techniques and methods of Scientometrics, because it is necessary to measure and analyze the theme proposed in this paper, individually, bibliometry is considered as the research method that through statistics and quantitative analysis are discovered trends and patterns of publications in a specific field of scientific literature are described [48].

In this way, a dataset was consolidated that was obtained from the results obtained from the bibliographic databases of Clarivate Web of Science (WoS) and Elsevier's Scopus. The search string used in these databases was “*((tourism OR tourist) AND (traceability OR tracing OR tracking OR trace OR flow OR movement OR pedestrian OR location OR following))*” to obtain the list of documents related to tourist traceability.

Subsequently, the dataset preprocessing was applied using the ScientoPy scientometric tool [49]. ScientoPy is a Python script tool specialized in the temporal analysis. The workflow steps of ScientoPy's scientometric analysis are: obtaining the dataset from the WoS and Scopus bibliographic databases, then the preprocessing where the document type

filters are applied, the correlation of the field labels, the simplification of the names of the authors, the omission of duplicate documents, the counting of citations, the analysis of the h-index and the extraction of countries and institutions, as a result of the preprocessing, generates a graph and a preprocessing summary. The third step is the analysis of data, through the main, specific topics, the search for occurrences and the analysis of trend themes. The last step corresponds to the visualization with several options: Graphs of timelines, bars, and parameters, and the word cloud. In turn, ScientoPy generates electronic sheets for each cluster analyzed, relating the detail of the related papers.

Figure 2.1 and Table 2.1 show the results of the datasets preprocessing. There are 4276 documents loaded. The processing has a filter of types of documents, those documents other than a conference paper, article, review, proceedings paper and article in press, were omitted, corresponding to 8.2% of documents loaded. Duplicate documents are also excluded, corresponding to 10% of the Scopus documents that were the second loaded dataset. In summary, 3534 documents were obtained as a result of the processing.

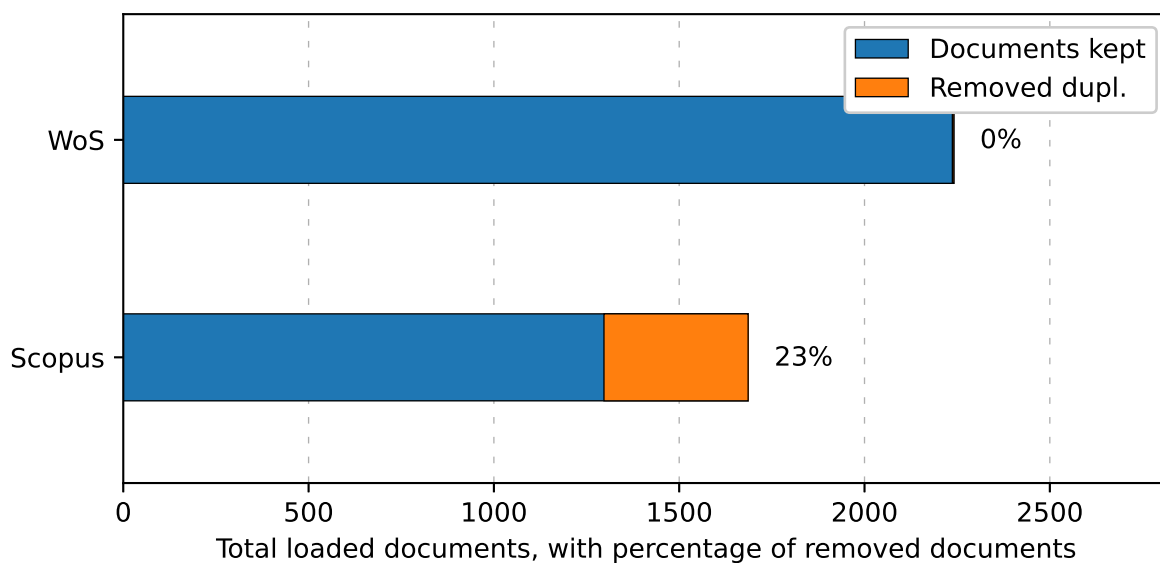


Fig. 2.1 The number of tourism traceability systems documents obtained from Scopus and WoS.

Table 2.1 ScientoPy preprocessing results of the tourism traceability systems dataset

Information	Number	Percentage
Loaded documents	4276	
Omitted documents by type	349	8.2%
Total documents after omitted documents removed	3927	
Loaded documents from WoS	2241	57.1%
Loaded documents from Scopus	1686	42.9%
Duplication removal statics		
Total duplicated documents found	393	10.0%
Removed duplicated documents from WoS	4	0.2%
Removed duplicated documents from Scopus	389	23.1%
Total documents after remove duplicates	3534	
Papers from WoS	2237	63.3%
Papers from Scopus	1297	36.7%

2.2 Tourist Traceability Top Topics

We consider traceability as the ability to access the information of an element in the course of a time-dependent system. The monitoring of the items allows identifying the origins and their characteristics for subsequent decision making. A traceability system is the set of disciplines of different natures that, coordinated with each other, allow the monitoring of the elements [50]. Traceability systems must track the history, application, and location of what is being considered [51].

Subsequently, the top keywords of the preprocessed documents are generated. Due to the number of documents, 1000 keywords are determined. Then these keywords, a filter of synonymy is applied, and in this way, in the first instance, the resulting top categories can be determined.

Figure 2.2 shows that the "Tourism" category presents the largest number of articles, logically because this term is part of the search chain. The category "Tourism Management" is a cluster that brings together research topics related to tourism management. The "Countries" category encompasses those investigations that focused on a particular country, region, or city. The category of "Tourist Experience" includes the investigations carried out in analyzing or evaluating a tourist's level of satisfaction when having consumed an experience. "Tourist factors" refer to the studies that evaluated those determinants of

tourist behavior. Studies related to tourist motivation are in the cluster with the same name. Studies related to lodging and hotels were grouped in the "Hospitality" cluster. More specific topics such as "Heritage" and "Sustainability" have their own cluster. However, the interest of our literature review research is focused on the remaining seven clusters: Destination, Data science, Location, Tourism technologies, Social media, Tracking, and Recommender Systems, because these clusters are closely related to tourism traceability.

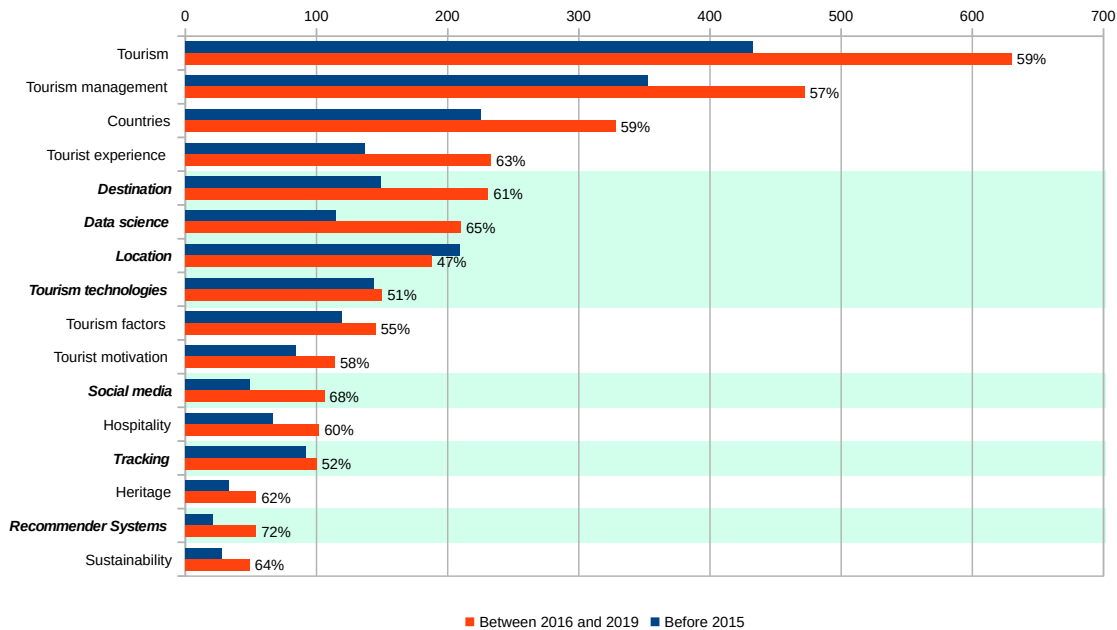


Fig. 2.2 Research trend in recent years of the top categories.

By calculating the number of annual documents on average (ADY), we can determine the absolute number of publications growth. The trend graphs in Figure 2.3 show the behavior after 2015 of the clusters of the tourist traceability systems. In the left part of the graph, we see the growth rate of the determined clusters' investigations. As of 2008, they have a very similar cumulative growth, except for the "Social Media" and "Recommender Systems" clusters, whose growth is a bit more dizzying. We can corroborate these curves on the right-side graph; these two clusters show the highest growth in recent years. Besides, we must consider the scientific interest for topics such as "Data Science," "Sustainability," "Tourist Experience," and "Heritage." The themes of "Tourism Management" and particular studies of some regions are the ones that accumulate more documents in the historical course, but not in the same proportion in recent periods. These parametric graphs show the growth in the number of documents and their relative growth.

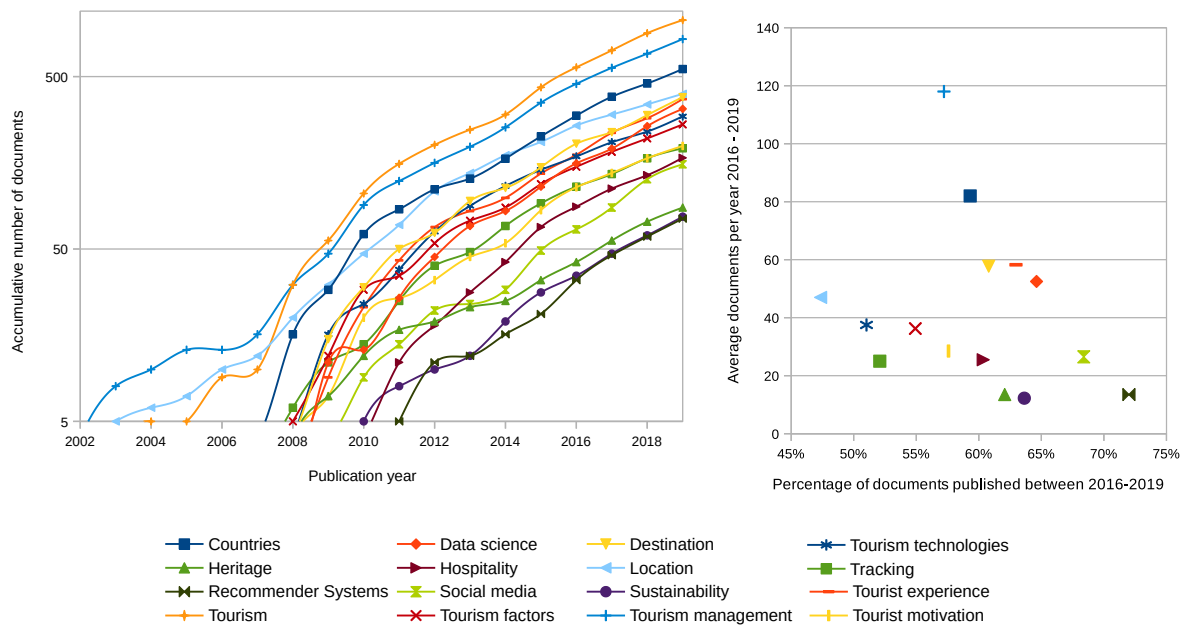


Fig. 2.3 Trend of publications of the top categories in recent years.

2.3 Location

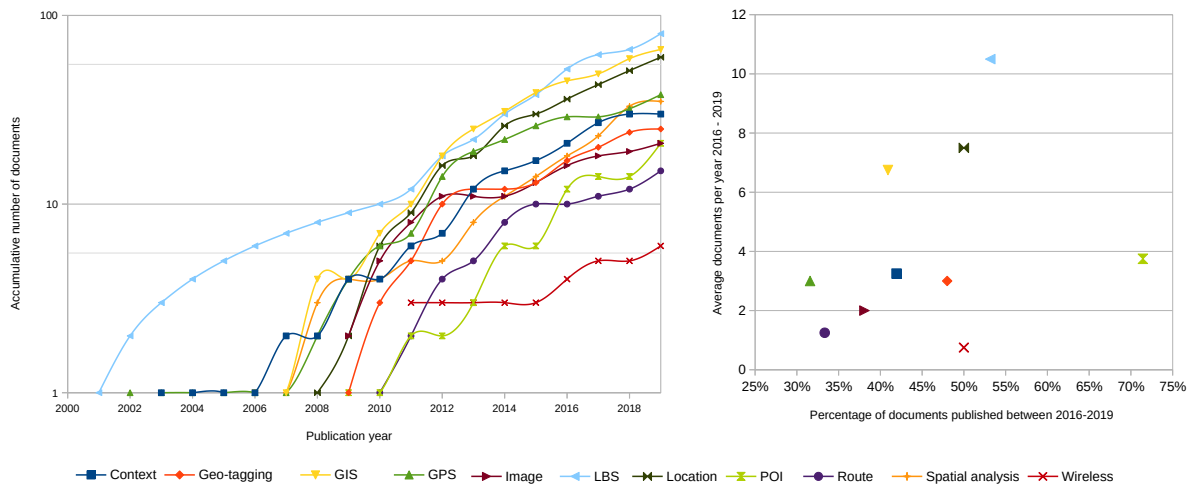


Fig. 2.4 Trend of publications of the top categories in recent years of the location cluster.

Researchers in the field of tourist traceability have focused their attention on different areas related to location. Such is the case of Geographic Information Systems (GIS), Global Positioning Systems (GPS), location-based services (LBS), cartography and geography,

geo-location, geo-tagging, maps, and spatial structure. Historically, GIS, GPS, and LBS are the topics with the most research documents related to tourism traceability. The beginning of these investigations coincides with the emergence and massification of the location technologies used in the tourism sector.

The location has been crucial in works on the analysis of tourist movements within a destination. [52, 53] identified different discrete patterns of movement through the use of GIS. These patterns are determined by factors such as territory, number of daily trips, number of detentions, tours, among others. In this way, we can know how the tourist consumes the destinations. Although GIS was beneficial, the authors highlight the complex nature of a tourist's movements and the technological limitations present at the time of developing their study. Similar work was done [54] when contrasting the behavior patterns of visitors for the first time and repeaters, using GPS and GIS. The authors concluded that first-time visitors tend to travel more extensively through the destination, while repeat visitors limit their travel to a smaller number of locations. Therefore, repeat tourists tend to use hotel services more. The study highlights the benefits of using advanced traceability technologies as they have the potential to transform tourism research by expanding its possibilities. In addition, [55] used GPS to assess the impact of weather on tourist movements, their findings have implications for destination management organizations (DMO), and for the tourist information staff because they allow us to understand the type of tourist activity that is weather sensitive or resistant.

Spatial data analysis is a set of spatial statistical techniques useful for describing and visualizing spatial distribution, detecting hot spots. When applying Moran I statistics in tourist flows, the significant spatial correlation is revealed [56]. The spatial analysis, supported by GIS, is also proposed for a market analysis for retailers who depend on tourism, to serve existing customers and to search more accurately for new customers [57]. This work is achieved using geocoding and geodemographic overlay techniques to develop tourist profiles. Subsequently, through screening and targeting techniques, new clients who meet a specific profile are determined. With a simulator of the individual spatial behaviors of tourists, the accommodation capacity of tourist centers can be assessed [58]. The microscopic simulation technique allows the analysis and prediction of travel patterns, location, cost, and status of tourist centers.

The Geographic Information Recovery System (GIR) [59] extracts and analyzes tourist information from photographs of online image collections (Flickr) by associating them with a city. Subsequently, use the Google Maps service to geolocate the recovered photos and analyze the referenced information. The system determines POIs and reconstructing the tourist routes.

The Location-Based Services (LBS) provide data according to the geographical location of the user through communications networks or positioning technologies. [60] developed a framework to provide information on Points of Interest (POI) near the tourist. The data is displayed through a GeoServer and implements Reverse Geocoding operations. Meanwhile, [61] uses LBS to support backpackers, with travel insurance and tourism planning services. These authors used kernel density estimation to calculate the critical points.

2.4 Global Position System (GPS)

GPS tracking of tourists has been applied in various situations. The authors [62–67] created a framework for monitoring tourists on their visit to destinations to determine their behavior and activity patterns. The tourist previously completes a profiling survey and uses a GPS device on the route. Besides, some research [68, 69] is leveraging GPS using Android apps and repository servers, such as Firebase, to provide smart tourist guides. Some researchers [70] avoid using surveys and using other techniques such as the CDR (Call Detail Records) that they obtain from cellular mobile phone companies, under data protection regulations.

Most studies use GPS to validate research findings from previous studies that did not use this technology. Researches on tourists' behavior need to combine approaches; the mapping needs a context that provides comprehensive information [71–73].

The literature review of tracking technologies from the first decade of the 21st century [74] shows three generations of research: methodological dimensions, temporality and spatiality data, and new data sources. Digital data have characteristics such as precision in temporality and spatiality, geographic coverage, and that can be complemented with other data sources.

Table 2.2 Summary of tourism papers that used GPS technology

Research	Year	Technology or Technique	Application
[52]	2008	GIS	Tourist behavior and movement patterns
[54]	2012	GIS	Tourist behavior and movement patterns
[55]	2015	GPS, GIS, Tracking Technologies	Tourist behavior and movement patterns
[56]	2013	GIS, Spatial data analysis	Tourist behavior and movement patterns
[57]	2008	GIS	Market analysis for tourism-dependent retailers
[59]	2012	Geographical Information Retrieval system, Geotagging, GIS, Flickr	Tourist information
[61]	2014	LBS, H-LBS	Backpackers
[62]	2018	GPS	Tourist behavior and movement patterns
[63]	2020	GPS	Tourist behavior and movement patterns
[64]	2019	GPS	Tourist behavior and movement patterns
[65]	2019	GPS, Time-geography (TG) framework	Tourist behavior and movement patterns
[68]	2019	Firebase, Android, GPS	Travel guide
[66]	2019	GPS, GIS	Tourist behavior and movement patterns
[70]	2020	GPS, CDR (Call detail records), GIS	Tourist behavior and movement patterns
[67]	2019	GPS	Marketing
[69]	2018	Android, Google Maps, GPS	Travel guide
[71]	2016	GPS	Spatial diffusion
[74]	2016	GPS, mobile, geocoded, bluetooth	Review of literature
[72]	2016	LBS, GPS, Google Maps	Travel guide
[73]	2018	GPS, GIS	Examining the terrain preferences in hazard conditions

2.5 Destination

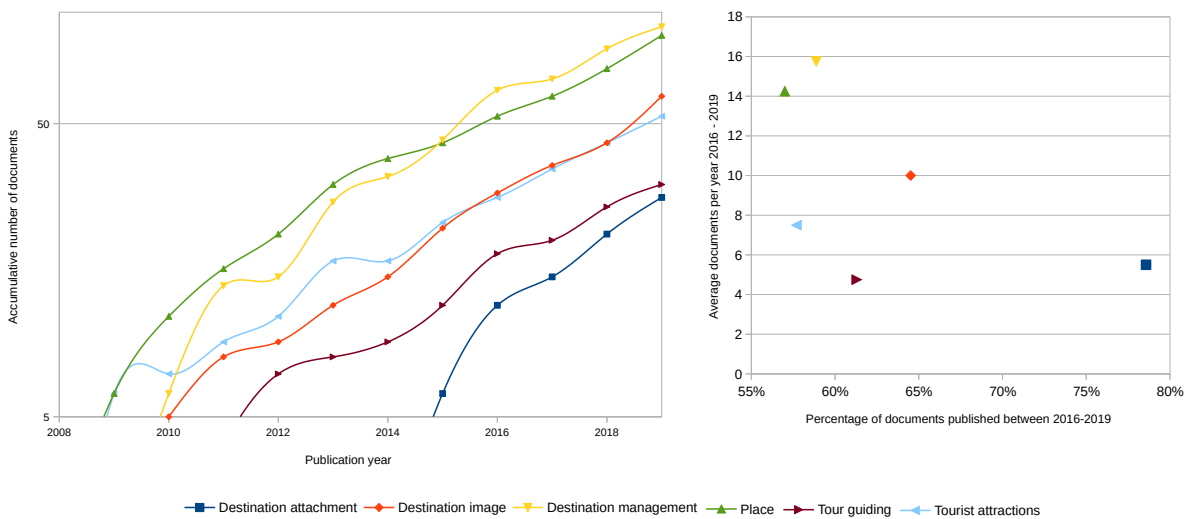


Fig. 2.5 Trend of publications of the top categories in recent years of the destination cluster.

Tourism traceability research related to the destination shows the higher volume of publications in areas such as destination image, hotels, hospitality, and context-awareness systems, these areas show an average of 75% on average of papers in recent years. Recently, new research themes related to the tourist destination have emerged, such as place attachment, walkability, sense of place, hotel chains, mega-events, place-making, and destination loyalty. Other areas with relative tendency are local development, place identity, destination branding, service quality, hotel industry, tourism promotion, and stakeholders.

2.6 Tourism technologies

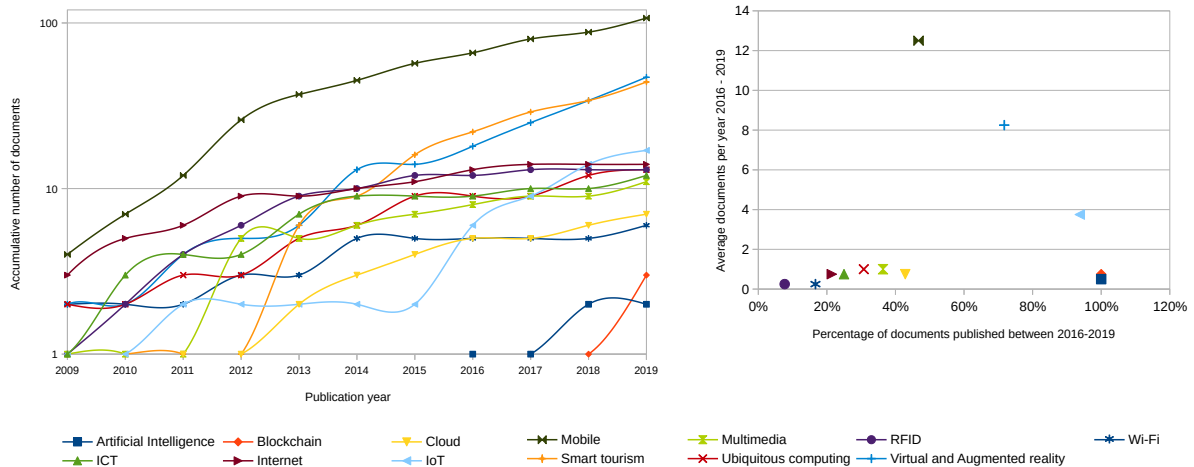


Fig. 2.6 Trend of publications of the top categories in recent years of the technologies cluster.

Research related to social networks that concern tourism traceability are the themes that bring together the most significant scientific flow within the ubiquitous computing cluster. The main social networks treated are Flickr, Foursquare, Twitter, Airbnb, Facebook, Instagram, and TripAdvisor. The second pervasive source is mobile computing. Research areas such as content analysis, semantic web, ubicomp, information retrieval, cloud computing, sensors, ontologies, and feelings analysis continue to be a trend. Within this category, research areas such as Web 2.0, accelerometer, and web services lose their validity.

2.7 Tourism domain ontologies

Some research about the semantic representation of the tourism domain uses information gathered from tourism websites for different applications. Xiang et al. [75] conclude that tourism websites can incorporate tools (such as reviews, tagging, and excavation) to allow travelers to interact directly with these sites. This way, the knowledge of travelers' perceptions and experiences can be collected and learned. Therefore, these tools offer promising avenues for tourist destination specialists to better understand and interact with potential visitors. Hence, the ontology is "the language of tourism" between the traveler and the industry. Online Travel Agencies (OTA) have communication channels with tourists to interact between them and the operators. Mainly these channels are based on reviews that tourists make about their experience; in general, OTA tag these reviews.

Concerning the knowledge domain of tourism, research such as Tribe and Liburd [76] "re-conceptualizes" its system taking into account three cores: Disciplinary knowledge, problem-centered knowledge, and value-based knowledge. The domain of TTS refers to the disciplinary knowledge of the ontology of this research. It denotes the importance of understanding that tourism is a multidisciplinary field and an extra-disciplinary one and considers the person, position, ideology, government, and global capital as elements of expertise. The "problem-centered knowledge" lies in the fact that Destination Management Organizations (DMO) need to have a knowledge base for decision making, especially statistical information obtained from the traceability of the tourist at the destination. The decision-making by the DMO enables the improvement of the destination infrastructure and the feedback of the tourism management system; in this way, we get value-based knowledge.

Mouhim et al. [77] highlight the importance of Knowledge Management (KM) in tourism: Share knowledge, facilitate the development of new products and services, develop the ability to learn, acquire tacit knowledge to transform it into explicit knowledge, satisfy customers and exploit the market. Based on [78], these researchers analyzed existing ontologies such as Harmonize Ontology [79], for the exchange of data between organizations; Mondeca Ontology [80] for profiling tourist and cultural objects, tourist packages, and multimedia content for tourism; and the OnTour project [81] that describes the domain of tourism focused on accommodation and activities. Seeing that none of these ontologies met the particular needs of their destination city, the researchers created their ontology (Moroccan Tourism Ontology), taking advantage of the thesaurus of the UNESCO and United Nations - World Tourism Organization (UNWTO). Its ontology has main classes: accommodation, transportation, attractions, activities, services, restaurants, and cultural heritage. As a study precedent to their OnTourism project, Prantner et al. [82] analyze,

in addition to the ontologies above, to the OTA Specification, the Tourism Ontology of the University of Karlsruhe, and the Travelling Ontologies EON and TAGA. They also review the main ontology management tools for the domain of tourism, identifying the following: DIP Ontology Management Suite, WSMT, WebOnto, and Ontolingua [83]. They complement the previous state-of-the-art because they feature a summary of ontologies in the travel industry, adding to the Comprehensive Ontology for the Tourism Industry, the LA_DMS project for destinations, the SWAP project, the Tiscover platform, and the Hi-Touch project, for the domain of Intra-European sustainable tourism. The ontology proposed in this document has as its domain the tourist traceability system in a specific destination. In contrast to the ontologies described above, we used Big Data analysis for the building of Ontology for Tourist Traceability (*OntoTouTra*). We collected data from ubiquitous computer sources, especially from social networks.

Subsequently, in the process of building a domain ontology for African tourism areas, Zhao et al. [84] reviewed new ontologies such as e-tourism ontology, Tourism ProtegeEsportOWL, the botanic ontology of the National Knowledge Infrastructure of the Chinese Academy of Sciences. In this way, they proposed a method of construction of ontologies in seven steps: Determine the field and the scope, examine existing ontologies, summarize essential concepts, define the classes and their hierarchy, define the attributes, define the properties, and finally, establish the individuals. We analyzed that the Big Data analytical methodology proposed by Erl et al. [4], in addition to contemplating the steps of the previous methods, is ideal for the collection and processing of large volumes of data at high transfer rates.

For unifying tourism terminology, we need a central authority that promotes standards for tourists and suppliers to understand tourism-related ontologies. Huang and Bian [85] recognize the UNWTO effort in defining the thesaurus about tourism and leisure activities but believe that it is not enough due to the complex character of tourist data. They propose their research to integrate both types of ontologies through the Formal Concept Analysis and Bayesian approaches. These approaches are mathematical tools for data analysis, knowledge representation, and information management, using triples with binary relations between concepts.

More recent studies, like Valls et al. [86], entrusted their research on word ontologies, such as WordNet (Miller, 1995), applying Clustering based on Ontologies, determining the motivations of tourists when visiting a destination. OnTraNetBD [87] also used WordNet for mapping the key concepts to build the ontology using the DERA (for Domain, Entity Classes, Relations, and Attributes) methodology [88] in six phases: Identify atomic concepts, analysis, synthesis, standardization, ordering, and formalization. From WordNet and Wikipedia were derived YAGO (for Yet Another Great Ontology) [89], which uses a logical

model, capable of representing n-ary relations maintaining compatibility with Resource Description Framework Schema (RDFS). In this sense, [90] developed a system that supports different types of document formats, including the essential structures of textual documents and native forms of the Web. In the paper, the authors compared the results of the semantic annotation approach with other popular methods (Armadillo, CERNO, CREAM, EVONTO, GoNTogle, KIM, MnM, Onto-Mat, and S-CREAM). Ontologies based on the word use relations between elements; for instance, Llorens et al. [91] named "term" to the words and established the relationships between the terms as the entity-relationship model of the UML diagrams in software engineering.

The tourism sector has highlighted the need to develop personalized applications using knowledge bases. Currently, researchers focus their interest on the development of applications based on ontologies. Such is the case of the scientometric review that we preliminarily carried out on frameworks of tourist recommendation systems [92] that is used heterogeneous data sources extracted from wearable devices, IoT, social networks, and ontologies. A specific application we find is the TRSO [93] recommendation system for tourists to know the attractions and the activities they can do. The recommender system uses collaborative filtering techniques based on information from attraction ontologies. Investigations like SocioOntoProcess [94] draw from social networks to build ontologies and take advantage of user interactions to develop the models, in this case, for consulting a consensual vocabulary. The ontology construction is collaborative through web tools, such as wikis.

SigTur/E-Destination [95] is a project, which from the knowledge management point of view, through a specific domain ontology, provides information on activities and guides aimed at the user and for employees. The system considers as much information as possible (demography, spatial, travel, motives, user stereotypes) to make the recommendations. We were also motivated to gather data from social networks, especially from OTA or electronic Word-of-Mouth (eWOM), because they tagged tourist reviews. Some OTA offer Application Programming Interfaces (API) to consult these reviews, but it is necessary to develop tools that can collect those public reviews for others. For this purpose, we create a Web Scraping tool.

From the perspective of the software industry, particularly the reuse of information, arose the RSHP meta-model [91], the authors looking for a general model capable of representing the information of software artifact, without dependence on their internal structure. They found that the data of all the artifacts form a representation of a particular domain. The authors concluded that the field could be created automatically by indexing the artifacts through a fundamental and simple idea: "the information is related facts." therefore, the central element of an artifact is the relationship. The semantics of RSHP qualifies

the existing relationship and its type; its components are Artefact, Term, Relationship, Information Element, and Property.

During the last decade, Shoval and Ahas [74] reviewed the literature on the use of tracking technologies for tourism, on average, 45 articles (40 percent of the articles published in the three leading tourism journals). This review found that tracking data occur in three generations: the first generation deals with methodological research and analyzes the potential of tracking data. The second generation is related to spatial and temporal data. The third generation is interested in new data sources. The researchers conclude that the movement of tourists has implications for infrastructure, transport, products, marketing, the commercial viability of the industry, and the management of the social, environmental, and cultural impact of the destination. They also detect the current research gaps in this area: a large amount of data for processing, personal data, and tourist data protection. Using new techniques is necessary to know the tourist traceability since some theorists think that the tourist can change the activity or behavior when being followed or studied.

Girardin et al. [96] proposed a challenge for social science research since large volumes of data from ubiquitous sources are available. With this data, we can understand the dynamics of the population and customize the services, among other essential activities for tourism management. They named the tourist tracks "digital footprints" that are of two types, active and passive. The passive traces are data left with the interaction of infrastructure, and the actives are the location data exposed by the users, especially in social networks. They worked with Flickr data (actives) and the call records of a telephone company (passives). The data used in Flickr is explicitly public data by the user. They carried out the process and the visualization of the large volumes of data through geo-visualization. Concerning data privacy, the authors handled the number of users instead of individual data. For this research, the expression "digital footprints" is similar to the data sources of ubiquitous computing, which are the input of the traceability system.

Mariani and Borghi [97] conducted a review of research literature in hospitality and tourism with Big Data and Business Intelligence for identifying future research and development gaps. They found that the research that applied analytical techniques is limited in scope and methodologies. Besides, conceptual frameworks are missing to identify critical business problems that link Business Intelligence and Big Data to tourism management. They evidenced epistemological dilemmas for the development of knowledge theories conducted by Big Data. They concluded with their study that further research on tourism should be stimulated and systematized by leveraging Big Data and Business Intelligence and providing information bases aimed at companies and stakeholders in tourism.

As a synthesis of this review of related work, Table 2.3 depicts the highlighted ontologies and their respective objective.

Table 2.3 Tourism domain ontologies found in the literature review

Ontology	Year	Purpose	TTS concepts covering?
Architectural ontology [98]	2018	e-tourism resources	No, it has an architectural domain.
OnTraNetBD [87]	2017	Uses WorNet for mapping key concepts	No, the ontology establishes the formal relationship between tourist attractions and other travel elements, but not the time-space causality of the tourist.
Ontology-Based Tourism Recommendation System [99]	2017	Travel Ontology	Partially. It defines a travel recommendation system based on ontologies but does not analyze tourists' routes in the destination.
Ontology-Based Human-Computer Cloud [100]	2017	Building ad hoc decision-support services	No, it describes various decision support scenarios in tourism in general but not specifically for TTS.
Dwipa Ontology III [101]	2017	Cultural parks, artists and monuments	No, it is limited to Point of Interest (POI).
TRSO [93]	2016	Recommender system for tourists	Partially. It determines the relationship of tourists with the context to suggest tourist information.
SigTur/E-Destination [102]	2011	Activities and guides	No, It provides a catalog of destination resources to offer personalized information to tourists.
Mondeca [82]	2011	Profiling tourist and cultural objects	Partially. Mondeca has a large number of concepts on tourism, but it is not freely available.
Moroccan Tourism [77]	2011	Ontology of this destination city	No, it is limited to presenting the importance of the knowledge domain in tourism.
University of Karlsruhe [82]	2007	OnTourism project for evaluating Semantic Web	No, they analyzed seven tourism ontologies and five management tools to create ontologies.
OnTour project [81]	2006	Accommodation and activities	No, it focused to e-tourism.
Harmonize Ontology [79]	2004	Exchange data between organizations	No, It was aimed at developing an interoperability platform for SMEs in the tourism sector.

In Table 2.3, we see that all ontologies meet a particular objective, which is why their domain of knowledge is well defined. We showed that none of the ontologies listed in this table have tourist traceability as their domain.

Chantre et al. [103] established two thematic cores of the movement of tourists and the tracking methodologies in the relationship of traceability and the tourist. In this sense, they considered tourist traceability as the set of actions, measures, and technical procedures to identify and record the activity of tourists in a given destination. For the above, to keep this record, it is necessary to build a spatio-temporal causality. Through a tourist traceability system, we gather information on the activities of interest to tourists, the most frequented POI, the timing of visits, tourist satisfaction with their experience, visitor profiling, and set a portfolio of tourist experiences, among others. In turn, a TTS allows decision-making by DMO, establishing Key Performance Indicators (KPI) that determine the level of service offered to improve destination management. With the above

considerations, it is essential to have a knowledge base of the [TTS](#) domain, with updated, accessible, actionable, and reliable data.

This study takes advantage of data from ubiquitous sources, especially from [OTA](#), because these satisfy the above requirements, especially tourist reviews. Furthermore, these allow identifying, among others, data on spatiality, temporality, satisfaction, feelings, preferences, and experiences. The analysis of this data is boosted through Link Data; for instance, with georeferenced data from tourist reviews, we reach more location levels, establishing a relationship between the review location and the hotel, destination, POI, or service reviewed. And so on, we move up the geographical level, passing through the state or region and reaching a particular country. Linking Data with Geonames provides complementary geographic information, which we did not obtain directly from the ubiquitous data source. Similarly, complementary temporal information is collected from linking data with the Time ontology.

The GeoNames ontology [104] allows adding semantic data to the World Wide Web. It has more than 11 million toponyms with a single URL (Resource Description Framework (RDF) web service). The ontology of GeoNames is available in Web Ontology Language (OWL) as a database dump and also as open data linked in RDF [105]. Geographic levels in GeoNames [106] vary according to the country; for example, Germany has six levels, France five, and Colombia four levels. Therefore, it was necessary to resort to national data providers, for the Colombian case, the National Administrative Department of Statistics (DANE) provides the DIVIPOLA system [107]. Thus, we can provide more data about the location of a person, hotel, or tourist attraction (POI).

The other aspect of the Spatio-temporal relationship of tourist traceability is based on temporal concepts; the [OntoTouTra](#) data link is the Time Ontology [108]. We took advantage of the vocabulary from this ontology to express the facts of relations between instants and intervals. We can establish temporal reference systems (time: `DateTimeDescription`), position in time (time: `TemporalPosition`), intervals (time: `DateTimeInterval`) and duration (time: `Duration` - time: `DurationDescription`).

[OntoTouTra](#) does not have a data link with any tourism management ontology. However, for its construction, Open Data repositories were taken into account by the International Open Data Charter [109]; for instance, we used Colombia's Open Data [110] and SITUR [111].

2.8 Tracking

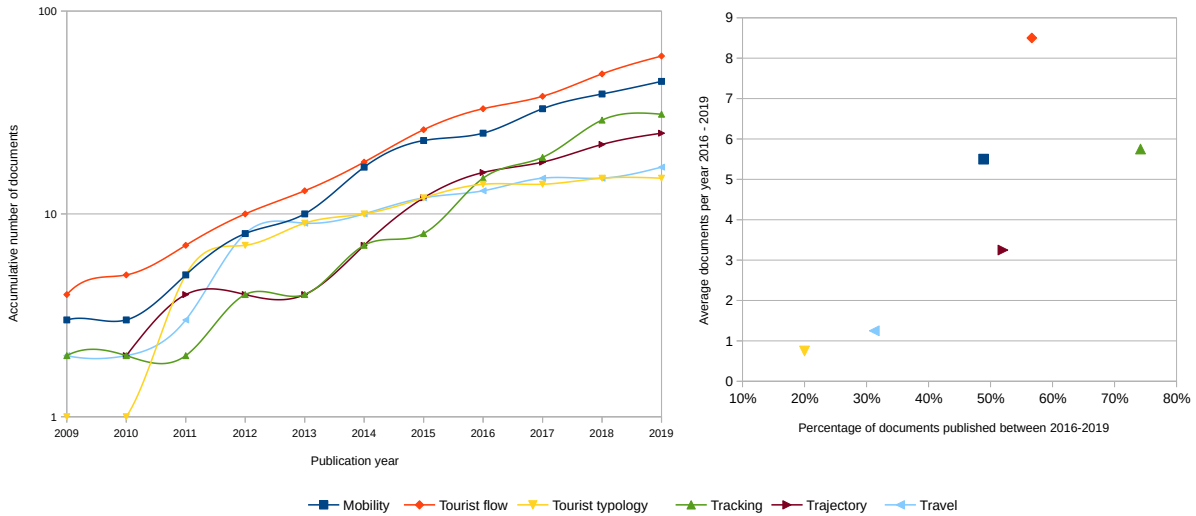


Fig. 2.7 Trend of publications of the top categories in recent years of the tracking cluster.

Mobility is the area that captures scientific attention within tourism traceability. The flow of tourists, accessibility, segmentation, route planning, space, and tourist guides are areas that maintain their tendency within this category. In recent years, new research themes have emerged, such as mobility patterns, tourist information, risks regarding data processing, monitoring, experience in the tourist flow, network analysis, and backpacking. Finally, mobile guides, pedestrian, movement patterns, and tourist transport have declined their research trend.

2.9 Data Science

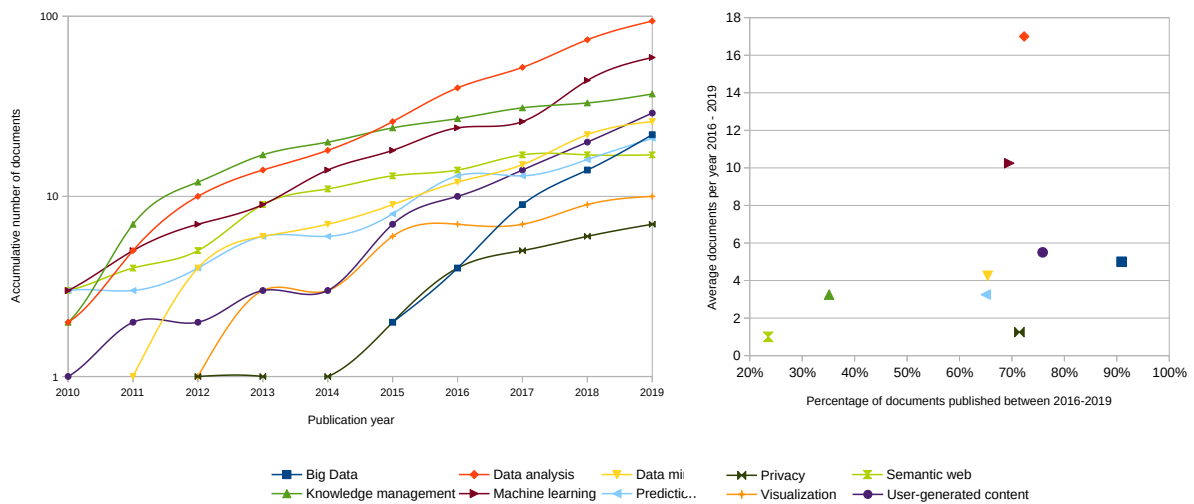


Fig. 2.8 Trend of publications of the top categories in recent years of the data science cluster.

Prediction and forecasting in the tourism context are the primary motivators for scientists to tackle machine learning. This category arises practically in the last decade, and the algorithms, classification methods, techniques, and data analysis within machine learning occupy the scientific interest to solve tourist traceability problems.

2.10 Cluster Mapping

For co-occurrence mapping, 444 terms related to traceability in tourism were taken into account. Thirteen clusters were obtained through the association strength method with a level of two and a resolution of 1 (see Figure 2.9). Distributed as follows:

- Cluster 1: These are the studies grouped to the right of the figure, corresponding to those investigations that focused on the use of technologies used, for example, GPS, Geo-tagged, LBS, ubiquitous computing, augmented reality, Big Data, social networks, among other.
- Cluster 2: It is located in the lower part of the figure and corresponds to the tourist experiences, motivation, and degree of satisfaction.
- Cluster 3: Corresponds to the destination and its management. It is counted in the upper left part of this figure.

- Cluster 4: Refers to accommodation systems. It is found in the lower right part of this figure.
- Cluster 5: It refers to regional tourist experiences and knowledge management. It is found in the lower central part of this figure.
- Cluster 6: Studies in the upper part of this figure are related to the flow and tracking of tourists and spatial patterns.
- Cluster 7: Refers to tourism policies and planning; it is found in the upper left part of this figure.
- Cluster 8: These are studies related to Big Data analytics, Intelligent Systems, and tourist demand; it is in the upper right part of this figure.
- Cluster 9: Studies on recommender systems, decision-making, and tourist behavior. It is located on the right side of the figure.
- Cluster 10: Literature review studies, ontological systems, semantic Web, and information retrieval are grouped. It can find this figure in the lower right corner.
- Cluster 11: It deals with socio-cultural aspects associated with tourism.
- Cluster 12: Studies related to the development of the destination such as sustainable development, marketing, and competitiveness.
- Cluster 13: Refers to opinion mining, review analysis, sentiment analysis, and prediction. It is in the upper right of this figure.

According to the nature of this research, the clusters that were not considered for analysis were 5, 7, 11, and 12.

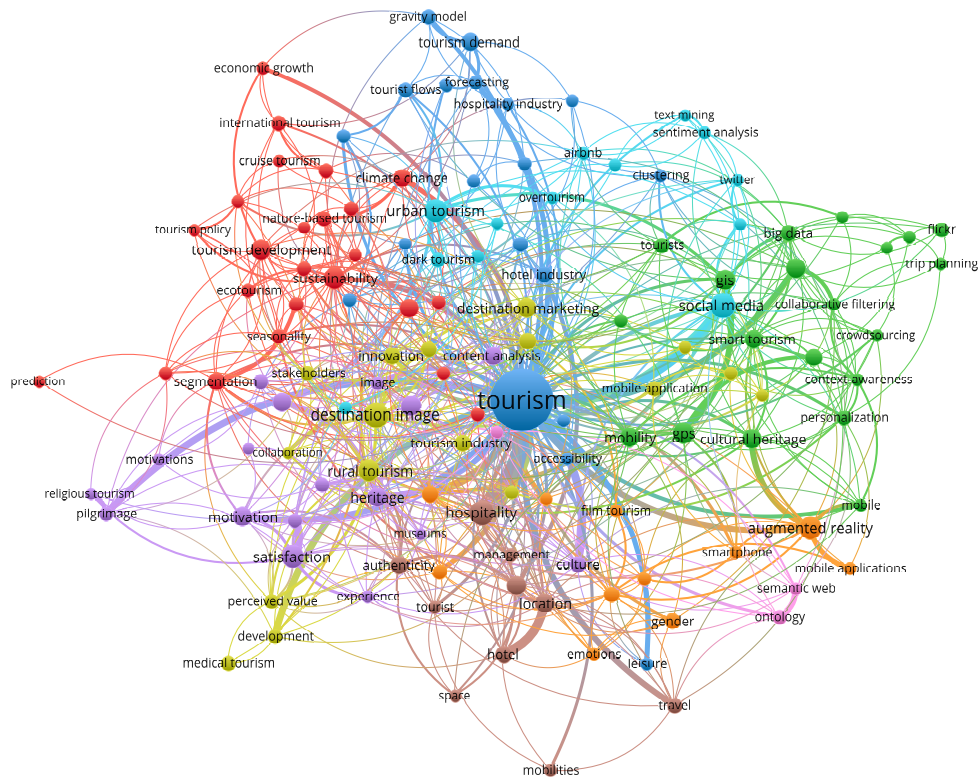


Fig. 2.9 Network Visualization.

Chapter 3

Tourist traceability conceptual framework

3.1 Tourism value chain

Based on the value chain concept proposed by Porter [112], Weiermair [1] built a value chain for tourism, as represented in Figure 3.1. The first activity of this chain is the provision of information and booking. From there, the model can vary according to the type of tourist service or product and the complexity of the trip. The actors involved in this value chain are suppliers, intermediaries, and tourists [2] (see Figure 3.2). For the success of the value chain, Zhang et al. [3] highlight the importance of understanding the features of tourism products and the tourism industry. In this sense, they identified six features of the tourist service, as shown in Figure 3.3.



Fig. 3.1 Tourism value chain (Based on [1])

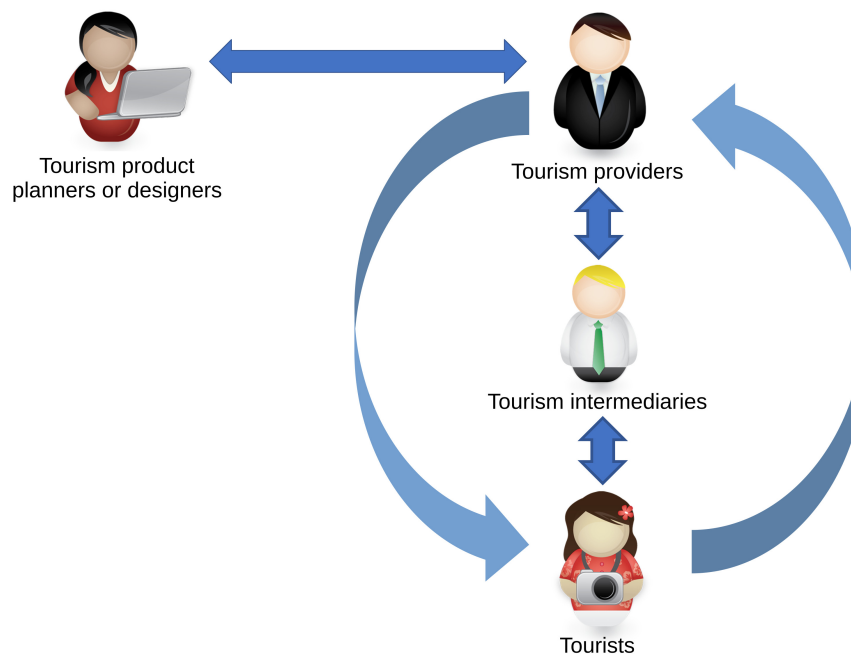


Fig. 3.2 Actors in the tourism value chain (Based on [2])

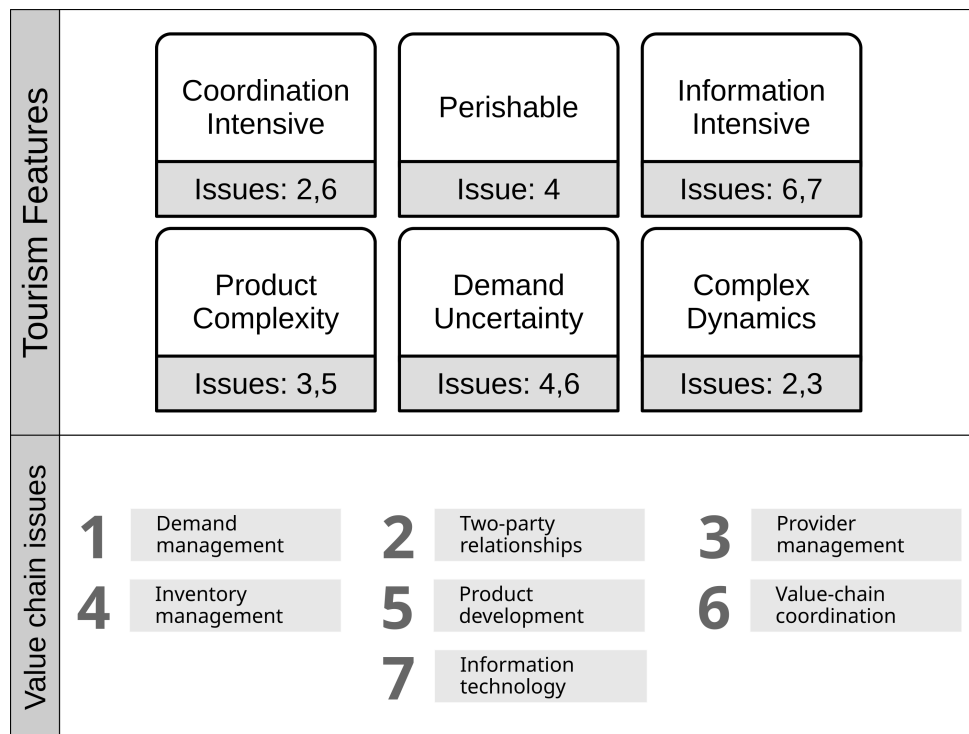


Fig. 3.3 Features of the tourist value chain (Based on [3])

Tourism is an intensive coordination industry where products and services are packaged to present a final tourism product. Tourism is also perishable, meaning it cannot be stored for future use. Besides, the tourist product is consumed in the destination, it cannot be examined before its purchase, and therefore, it depends on its presentation and the information provided. Hence, tourism products are complex because they have different service components and present uncertain demand and intense competition between suppliers.

3.2 Methodology for the construction of the conceptual framework

According to the state-of-the-art results, we recognize the recent interest of researchers towards tourism traceability. As this research aims to analyze some tourism traceability alternatives, we defined a conceptual framework, clarified concepts, and determined reference points to evaluate the alternatives weighed. According to [113], we tried a conceptual framework as the network of interconnected concepts to understand a phenomenon. In this case, the thematic domain is tourism traceability, not understood as a simple collection of concepts but the representative role of each concept within the domain.

The methodology for the analysis of a conceptual framework [113] comprises eight phases:

1. **Mapping of selected data sources. We analyzed the following sources related to traceability:**
 - According to ISO 9000 [114] and ISO 9001 [115], the definition of the traceability term.
 - The clause of ISO 9001:2015 for traceability [115].
 - Bibliometric analysis of the scientific databases Scopus and WoS (see Chapter 2)
 - Papers on secondary studies on tourist traceability [103].
2. **Extensive reading and categorization of selected data.** We analyzed these aspects of categorizing the information found in the data sources: Traceability, tourism traceability, components of tourism traceability, and management of tourism traceability.
3. **Identification and naming of concepts.** We looked for conceptual frameworks for traceability of different domains [116, 117] to obtain a concept discovery guide and their connection.
4. **Deconstruction and categorization of concepts.** We identified each concept's attributes, features, roles, and categories according to its ontological, epistemological, and methodological function.
5. **Integration of concepts.** Concepts with similarities were grouped into a new concept in order to reduce their number.
6. **Synthesis, resynthesis, and making it all make sense:** It is an iterative synthesis process to recognize that the framework makes sense by verifying the conclusions.
7. **Validation of the conceptual framework:** Chapter 5 builds an [OntoTouTra](#) from this conceptual framework and presents the validation within the tourism traceability domain by testing a use case.
8. **Rethinking the conceptual framework:** This is the main phase of the conceptual framework since we identified reference points to evaluate some tourism traceability alternatives as to the primary purpose of this research work.

3.3 Conceptual framework

As a result of applying the previous methodology, we got the concepts and their relationships that facilitated the preliminary identification of the domain of tourism traceability (see Figure 3.4).

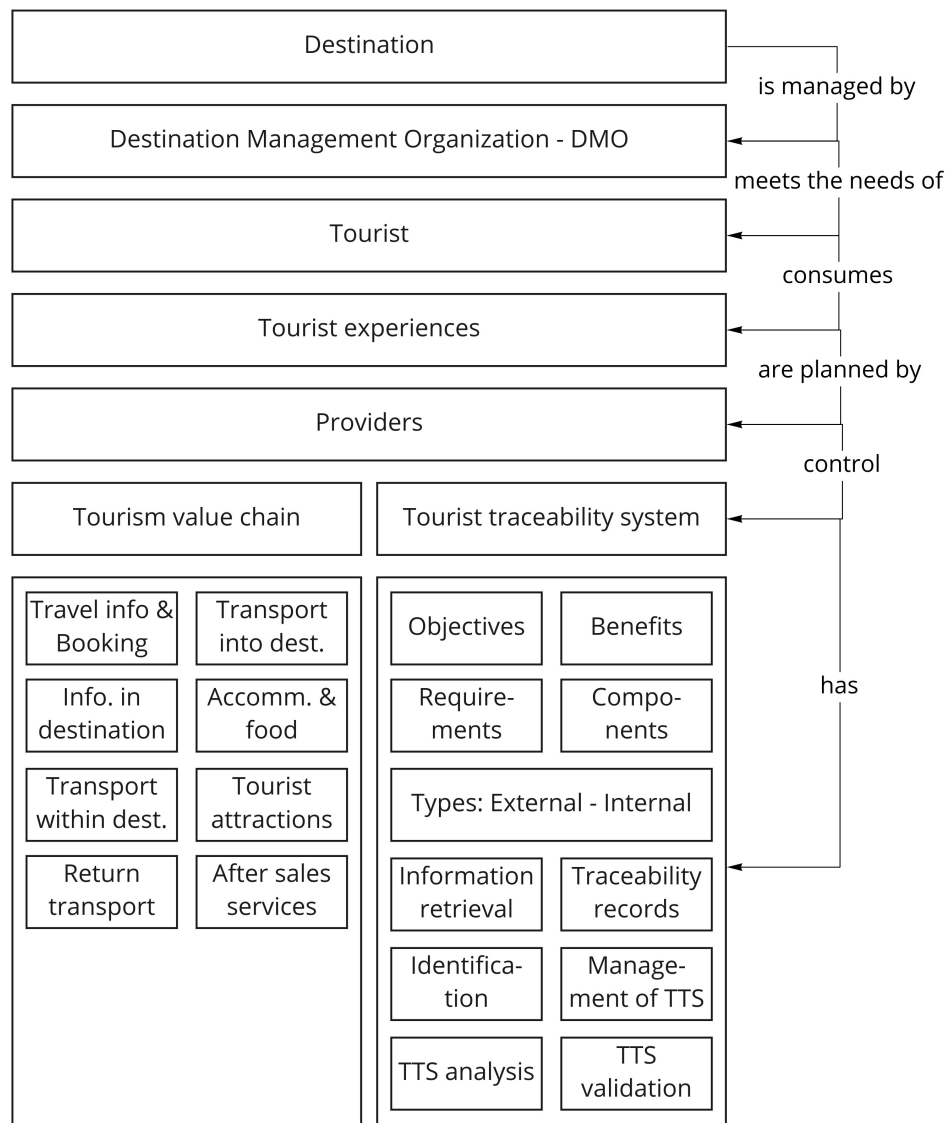


Fig. 3.4 Conceptual framework for TTS domain

We proposed this conceptual framework to clarify the concepts and their relationships within the tourism traceability initially proposed by Chantre et al. [103]. They analyzed the literature on tracking, tracing, and the flow of movement of tourists within the destination. This review detected the lack of capacity to trace the value chain processes and the

tourism supply chain. In order to maximize the tourist experiences in the destination, it is necessary to apply traceability in the stages of these chains. As a result, we will obtain the optimization of the tourist service and the improvement of the infrastructure in the destination.

Below we present the main concepts and their relationships in this conceptual framework around tourist traceability.

3.4 Background and overview

3.4.1 Traceability

It is the ability to trace an object's history, application, or location [114]. When considering a product or service, traceability may be related to its origin, processing history, distribution, and location of the service or product. A more specific concept is found in [116, 117]: "traceability is the ability to trace, track and identify units of the product only through a defined supply chain or production operation."

In tourism, we apply traceability to the value or supply chains. The product or service units refer to the touring of tourists within the destination, visiting its **POI**. In Figures 3.1 and 3.3, we see the stages of the value chain and the possible issues that may arise in these stages. Tourist traceability allows us to detect, follow, review the history, feedback, solve, and improve the service in each of these stages. The stakeholders involved in managing tourist traceability are the **DMO**, specifically the providers and managers of the service. Tourist traceability should not be considered as the individual monitoring of a tourist. It is a quality management system for the continuous improvement of the tourist service. The monitoring of the touring is done to the visitor groups.

3.4.2 Tourist traceability

It is the ability to trace, track and identify tourism services and products through the value chain while tourists consume the experiences at the destination.

3.4.3 Tourist Traceability System **TTS**

It is a set of interrelated components that allows recording tourists' history, application, or location. The **TTS** collects data from the touring. In this way, **TTS** generates essential information for decision-making on destination management, detecting the stage to improve within the value chain.

3.4.4 Tourist traceability considerations

- Although in the literature review, we did not find that tourist traceability is a legal requirement, as is the case with traceability in other contexts such as food, health, among others, it is an essential factor for continuous improvement in service provision.
- Tourist traceability can be considered a management tool, which helps us ensure that the tourism service or product improves continuously and, therefore, maximizes the satisfaction of the tourist experiences offered in the destination. It also allows taking the necessary actions if any component of the system presents any non-conforming.
- Traceability identifies the path from which the service originated and who provided it. It is a series of records of the stages of the value chain process while tourists visit the POI of the destination.
- Traceability is a widely used term, of which there are many different applications [116, 117]. There is no universally accepted definition of the traceability system.
- Unlike other traceability contexts, the tourism product or service records the activities of tourists in groups within a destination. Therefore, it is imperative to consider the regulations for the treatment of personal data and similar legal provisions to prevent the violation of people's privacy.
- A TTS should be considered a management system for decision-making by DMO. Typically, these systems should generate information related to the performance and monitoring of tourism service providers and service quality.

3.4.5 Aims of the implementation of tourism traceability

- Ensure the continuous improvement of the management of the tourist destination.
- Provide a rapid response to incidents detected at any stage of the value chain.
- Enable access to information about the components of the tourism service.
- Provide support information to respond to the suggestions of tourists regarding their experiences.
- Improve the provision of the tourism service.
- Adapt the infrastructure of the destination according to the tourist's requirements.
- Identify the actors responsible for non-conformities in the tourism value chain.

3.4.6 Benefits of tourist traceability

- Support for continuous improvement in the tourism service and quality objectives, and satisfy the needs and expectations of the tourist.
- Fulfil with the guidelines of the quality standards of the provision of the tourism service.
- Transmission of the information to the stakeholders of the tourism service.
- Satisfaction of the tourist experience.
- Support for decision-making by [DMO](#).

3.4.7 Components of tourist traceability

TTS has three main components:

- Traceability of the tourism service provider
- Traceability of the tourism service
- Traceability of tourists

3.4.8 TTS information categories

- Internal traceability: Relates the history of the tourist service with its provision. For example, the identification and tracking that is done, where the service comes from, when and how the service was provided, and the identification of the infrastructure used to provide the service.
- External traceability: In turn, it is of two types. First is the external traceability of the service to the provider; an example is the follow-up of tourist experiences by evaluating the providers' performance. Second is the external traceability of the service to the tourist, and an instance is the feedback of tourist reviews.

3.4.9 Registry keeping

As we can see in Figure 3.1, the value chain of the tourism service is a sequential series of separate operations (from preliminary information on the trip to after-sales services). Tourist traceability is done from tourists' point of view when they interact directly in each of the stages of the tourism value chain.

The **TTS** records the information on the service giving by the providers responsible for each of the stages of this value chain. By linking the data resulting from each step of the chain, we achieve the traceability of the chain in its entirety.

3.5 External tourist traceability

3.5.1 Applicable businesses

Tourist traceability can be applied to any **DMO**, regardless of the size of the destination, the providers, or the type of tourist experiences offered. However, it is necessary that the **DMO** is consolidated and that the value chain has been identified and is operational.

3.5.2 Requirements around the tourist

It is required that providers and **DMO** have fully identified the types of tourists visiting the destination, as needed for the **TTS**. In addition to identifying providers, it is necessary to identify the services and tourist experiences in the destination.

3.6 Internal tourist traceability

3.6.1 Tourism service traceability

Internal traceability operations can be:

- **Transfer:** Services are transferred directly as a unit from one stage of the process to the other. For instance, the booking transfers to the accommodation.
- **Joining:** In one stage of the service, several services are combined. For example, touring combines various services such as transportation between **POI**, entrance to attractions, food, tour guide, among others.
- **Splitting:** A service unit is divided into different processes. For example, in the information service stage at the destination, it can be used for other services, such as transportation, accommodation, food, attractions, among others.

3.6.2 Batch identification

The batch group together processes to form a tourism service. Each batch is identifiable and must be separated by space-time causality. Batches depend on the nature of the service, and their monitoring contributes to traceability.

3.7 Retrieval of traceability information

The moment of recovery of the traceability data depends on the complexity of the tourism service and the number of tourists in the destination.

3.7.1 Timeframes

The appropriate period is related to the features of the tourism service, the complexity of the tourism ecosystem, the features of the destination, the profile of tourists, seasonality, the maturity and experience of the providers, and the requirements of the [DMO](#) to retrieve traceability information.

3.8 Product units of tourist traceability

At each stage of the value chain, we identify units of tourism service or product. Each specific unit has a unique identifier or code that detects a service failure in the stage or when we want to improve it.

The flow of tourists determines the units in the stages of the value chain. Each unit depends on the nature of the service provided at each stage. Units can be presented that group several elements of the service, which we call batches. The traceability that is applied to the units ends in a time-space relationship. The operations on the units are closely related in such a way that they constitute processes.

Performance measurements can be applied to unit operations to determine possible failures or improvements in the tourism value chain. Thus, in the first stage of the value chain (see [Figure 3.1](#)), we can identify units such as the number of tourist reserves depending on the operations to be traced. A performance index (KPI) is the effectiveness of the reservation service, which would be calculated as [Equation 3.1](#):

$$KPI_{E_{Booking}} = \frac{E_{Attended}}{E_{Total}} * 100 \quad (3.1)$$

where:

$E_{Attended}$ = Number of bookings attended;

E_{Total} = Total number of bookings;

In the same example, another case of traceability application would occur in the event of a non-conformity. The tourist made his/her reservation, but the provider could not comply with the service (see Figure 3.5).

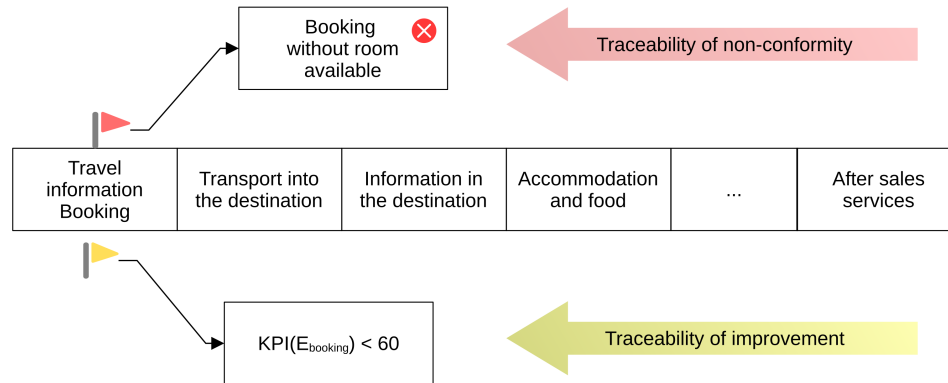


Fig. 3.5 Example of two scenarios of tourist traceability of booking service in the first stage of value chain

3.9 Identification codes and marks on TTS

At each stage of the value chain of a tourism system, we find value chains corresponding to that stage. Therefore, the units of tourism service or product vary between stages. In the same way, the identification systems of these units are different. An identification coding system is significant for the tourism traceability system because it recognizes the product by linking operations between stages or sub-stages. Before activating a traceability alarm, we will be able to trace the service in time and space and thus determine the actions to be carried out through the identification system.

The identification of the tourism service is based on assigning an identifier to a unit of service or product. Hence, this identification is added to or accompanies the unit through the chains of operations through an identification carrier. The identification can be a set of numeric or character digits or both. Depending on the degree of formalization of the value chain of each stage, the identification configuration can be standardized, and its system is open. Ideally, this identification system should include metadata such as the date and location of service provision, its type, among others. The identification system must allow human or machine reading through commonly used coding systems such as barcodes, QR codes, RFID, WSN, and NFC.

For instance, in stage 2 of the value chain (see Figure 3.1), transportation into the destination, in turn, involves value sub-chains, depending on the needs or type of service. In the case of air transport, the IATA has established the PNR (Passenger Name Record) as the unique identifier of the air ticket service [118]. The PNR code consists of six or ten alphanumeric characters. It is a digital certifier that allows multiple operations such as ticket purchase, check-in, baggage reception and claim, and even in some cases, booking.

3.10 Management of the TTS information

How traceability information is recorded depends on the nature of the tourist service and its operations. This registration procedure is essential because the information retrieval depends on it (see Section 3.6.1). TTS records can be based on paper or electronic devices. Regardless of these storage methods, it is required to guarantee the organization and retention conditions to allow easy retrieval. In this regard, to estimate the retention time of the information, it is necessary to consider the nature and features of the service and the traceability requirements. However, the more information is kept, the easier and faster it will be to identify the affected service, mitigate risks, save time and money [119]. It is advisable to keep the traceability information of the previous, current, and successive value chain stages (one step back and one step forward), considering that the traceability information must be available on demand.

3.11 TTS analysis

The analysis starts from a detailed review of the value chain and determining the stage to be traced. The traceability information of the contiguous stages (previous and next) is obtained for each stage, selecting the traceability units and their identification. Subsequently, the chain of operations affected is determined. When detecting the issue in any of these stages, the approach is applied one step backward or one step forward to search of the source of the issue or advance to determine the impact (one step back and one step forward). The analysis determines the identification details, the recovery of the registered traceability information, and the operations involved.

For example, consider the following review of a tourist (Table 3.1), obtained from an OTA:

Table 3.1 Example of TTS metadata

date	14 April 2018
location	987654321
hotel	123456
user	xxxxx
country	UK
score	8.3
review	"...the showers are solar so not always warm and when we went the electric shower was down..."

The table 3.2 depicts the traceability analysis ("Accommodation stage"):

Table 3.2 Traceability analysis - One Step Back

PN	Process step	Identification read	Recorded data	Note
3	Reviewer feedback	location, hotel, OTA	Review	N/A
2	Clean and tidy up hotel rooms	Hotel housekeeper, date	Cleaning report	N/A
1	Hotel Facility Maintenance	hotel, date	What the room is Last Maintenance Who was the maintenance worker Last maintenance date Maintenance report	Maintenance of the electrical system of the showers

3.12 TTS validation

Considerations for an effective traceability system include [120]:

- Stakeholders
- The service and processing locations
- The services that a provider uses or creates
- Service units
- Identification of the units

Many DMOs verify the effectiveness of traceability based on tracking activities, measuring the ability to supply the information within defined time frames. A validation analysis method can be "input / output." Where the identification, service process, result, and tourist feedback (reviews) are checked. In each operation, compliance with two required components is verified: The identifier and the service history record.

The Requirements Traceability Matrix (RTM) is a method of validation of traceability systems used mainly in the software industry. RTM is a document that links the requirements during the validation processes. The matrix crosses the test cases with the requirements, thus ensuring compliance with the needs and, therefore, the traceability system's effectiveness. The Table 3.3 depicts a RTM example:

Table 3.3 RTM Example

Test case	Requirements					
	Customer satisfaction	Fast response time	Profitable	Safe	Sustainable	Reliable
Booking						
Travel info						
...						
After sales services						

In the matrix of Table 3.3, six requirements were defined to track in each test case described in the first column. Check the corresponding box if the traceability of the requirement in that test case was met. Subsequently, the cells where some issues were detected are analyzed, and the traceability method is applied to detect the source and impact of the issue (one step back and one step forward).

In a non-conformity of the tourist service in one of the value chain stages, the DMO controls (make a decision) the TTS immediately. The decision tree in Figure 3.6 outlines the process to follow.

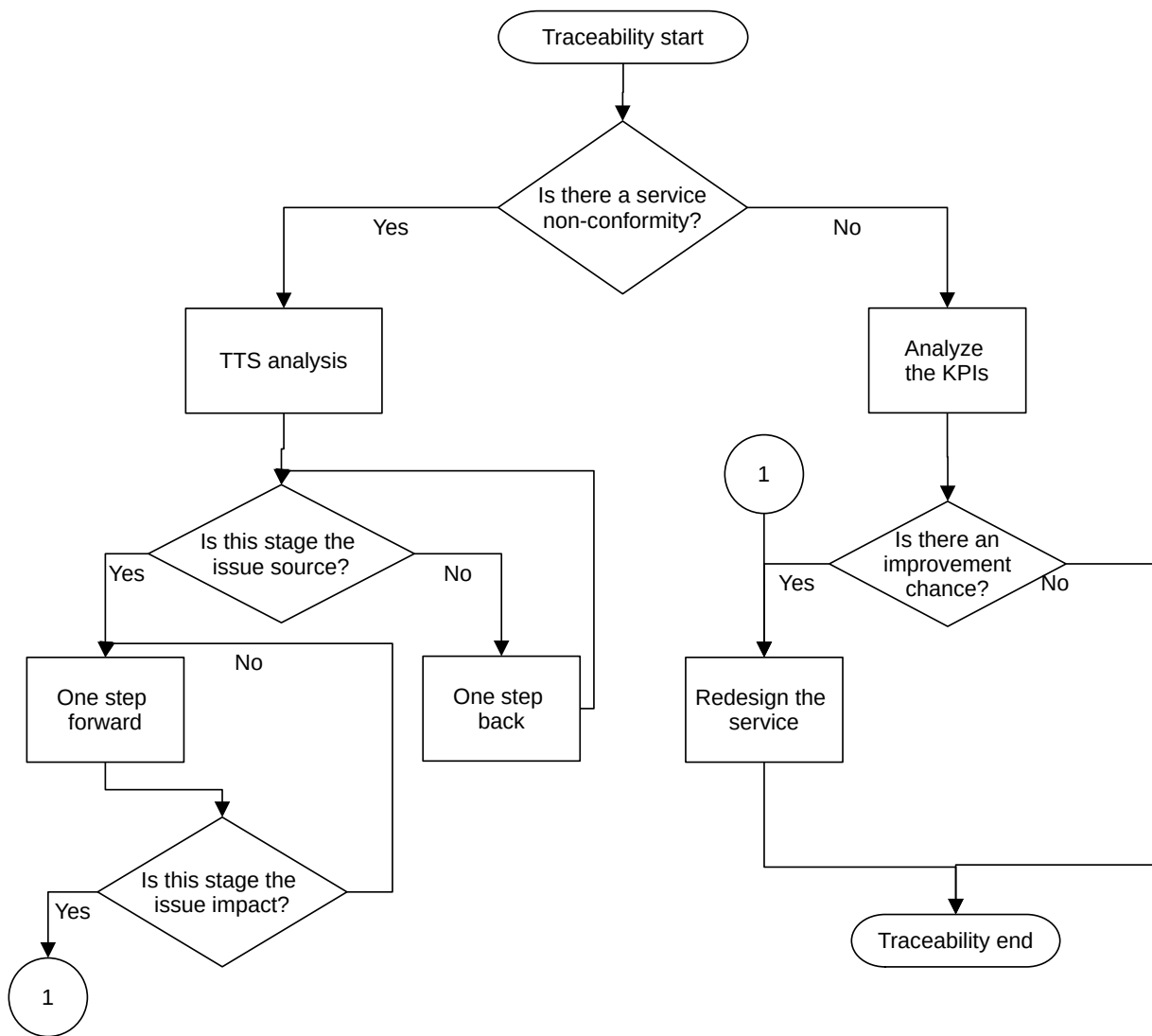


Fig. 3.6 TTS making decision

Chapter 4

Tourist traceability alternatives

4.1 Related work

Research on tourist traceability is relatively recent; studies on TTS modeling are scarce. However, to establish an analysis of TTS alternatives, we carried out a systematic mapping that describes the nature of research in tracking, tracing, or trajectory analysis in the tourism sector. These areas show degrees of affinity with tourist traceability. In this way, we applied the search string "TITLE-ABS-KEY ((tourist OR tourism) AND (tracking OR tracing OR trajectory)) AND (LIMIT-TO (SUBJAREA,"SOCI") OR LIMIT-TO (SUBJAREA,"BUSI") OR LIMIT-TO (SUBJAREA,"COMP") OR LIMIT-TO (SUBJAREA,"ENGI") OR LIMIT-TO (SUBJAREA,"DECI") OR LIMIT-TO (SUBJAREA,"MULT") OR EXCLUDE (SUBJAREA,"EART") OR EXCLUDE (SUBJAREA,"ARTS") OR EXCLUDE (SUBJAREA,"MATH") OR EXCLUDE (SUBJAREA,"ENER") OR EXCLUDE (SUBJAREA,"PHYS") OR EXCLUDE (SUBJAREA,"AGRI") OR EXCLUDE (SUBJAREA,"MEDI") OR EXCLUDE (SUBJAREA,"MATE") OR EXCLUDE (SUBJAREA,"PSYC") OR EXCLUDE (SUBJAREA,"BIOC") OR EXCLUDE (SUBJAREA,"CENG") OR EXCLUDE (SUBJAREA,"CHEM") OR EXCLUDE (SUBJAREA,"HEAL") OR EXCLUDE (SUBJAREA,"NEUR") OR EXCLUDE (SUBJAREA,"NURS") OR EXCLUDE (SUBJAREA,"PHAR"))" in the Scopus database, obtaining 1009 results. Subsequently, the relevant papers of those studies that proposed models similar to TTS were identified. This identification was supported with the literature review on the analysis of trajectories in tourism [121, 122] and with the conceptual framework on TTS (see Chapter 3).

The co-occurrence mapping is analyzed to identify the themes related to the transversal axis of the model for tracking the tourist. For this purpose, the bibliographic data extracted from Scopus were used to generate a network map with the VOSViewer tool [123]. Initially, the author's keyword and co-occurrence map was created with a complete counting method. The network map formed ten clusters from selecting 2998 keywords, of which 56 were found at the threshold. For each of the 56 keywords, the tool calculated the total strength of the co-occurrence links with other keywords. The merged network represented the evolution of the themes over time (2012 to 2020), showing the most significant traces of the related

research documents (see Figure 4.1). Each point represents a node in the network, and the lines connecting the nodes are co-occurrence links. Three clusters were selected that show homogeneity with the thematic category.

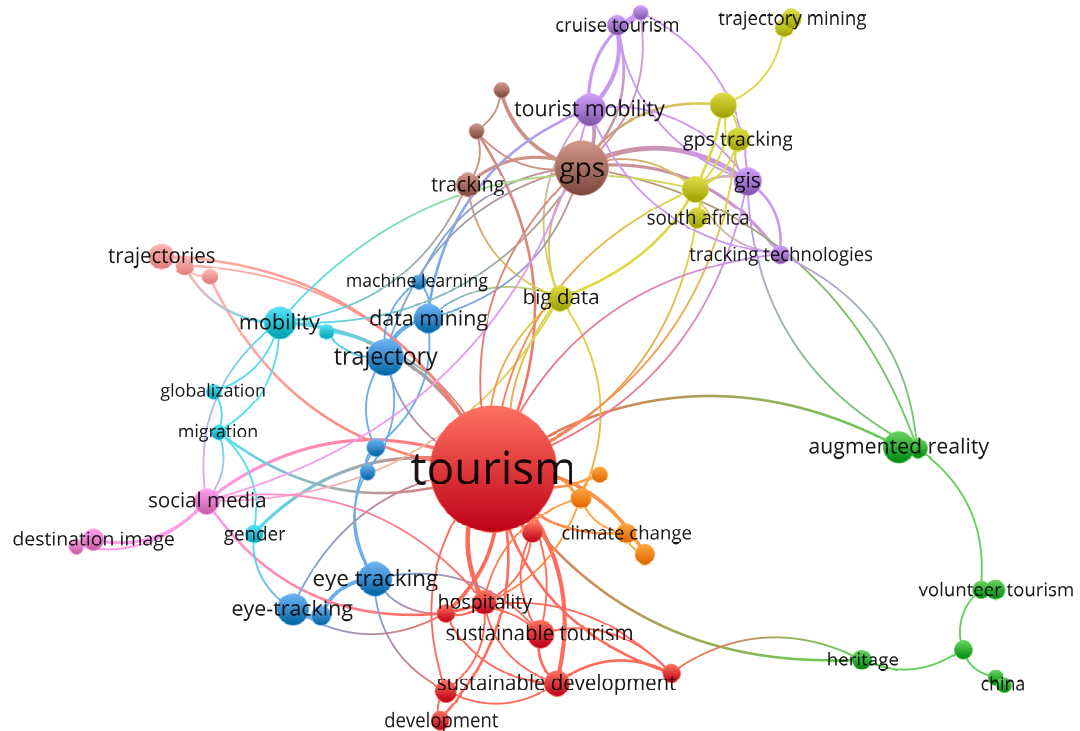


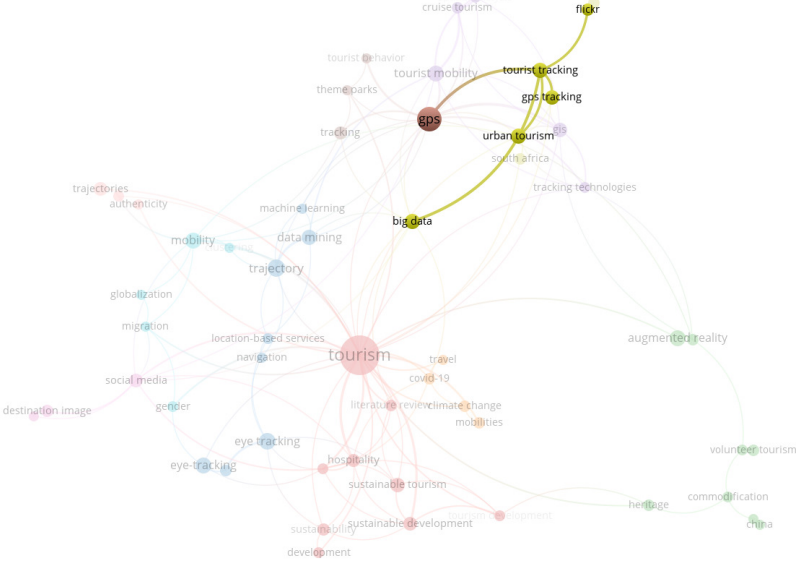
Fig. 4.1 Mapping of co-occurrence networks of author keywords related to the transversal axis of tourist tracking models. Displays ten colored clusters made up of nodes identified by labels. The grouping of related documents defines the size of the nodes and the width of the lines between the nodes.

Apart from tourism, which was the domain of the search, the keywords with the most significant link strength were: GPS, tourist mobility, GIS, social media, mobility, trajectory, Big Data, and tracking technologies; with strength indices of 35, 18, 17, 14, 13, 11, 10, and 9, respectively.

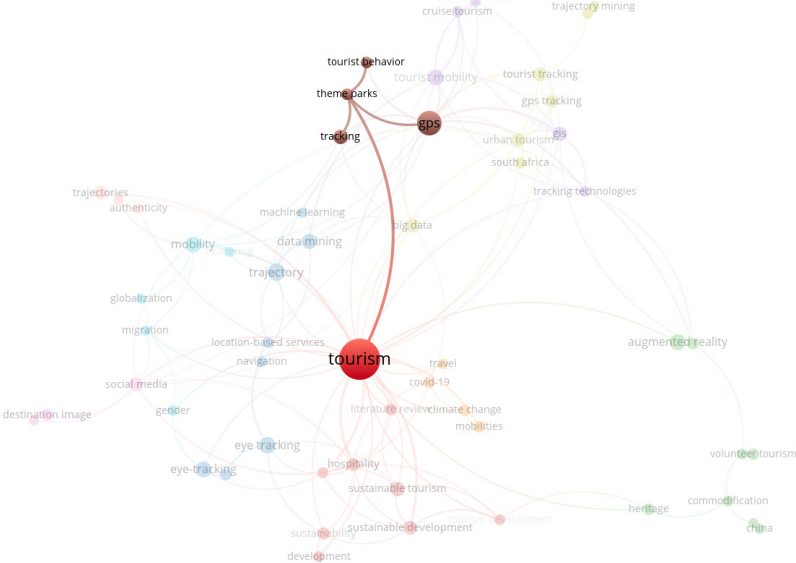
The three related clusters were:

- Figure 4.2a shows the configuration of the first cluster that focuses on tracking tourists using GPS technologies supported by Big Data in urban environments.
- Figure 4.3a represents the second cluster on the analysis of tourist mobility with GPS and tracking technologies, supported by data mining techniques and Big Data for data from social networks. This cluster had the most significant trend, as can be seen in Figure 4.3b.

- Figure 4.2b describes the relationship of cluster 3 focused on tracking tourists to analyze their behavior with GPS technologies in environments such as theme parks.



(a)



(b)

Fig. 4.2 Clusters of tourist tracking using GPS technologies: a) supported by Big Data; b) for the behavior’s analysis.

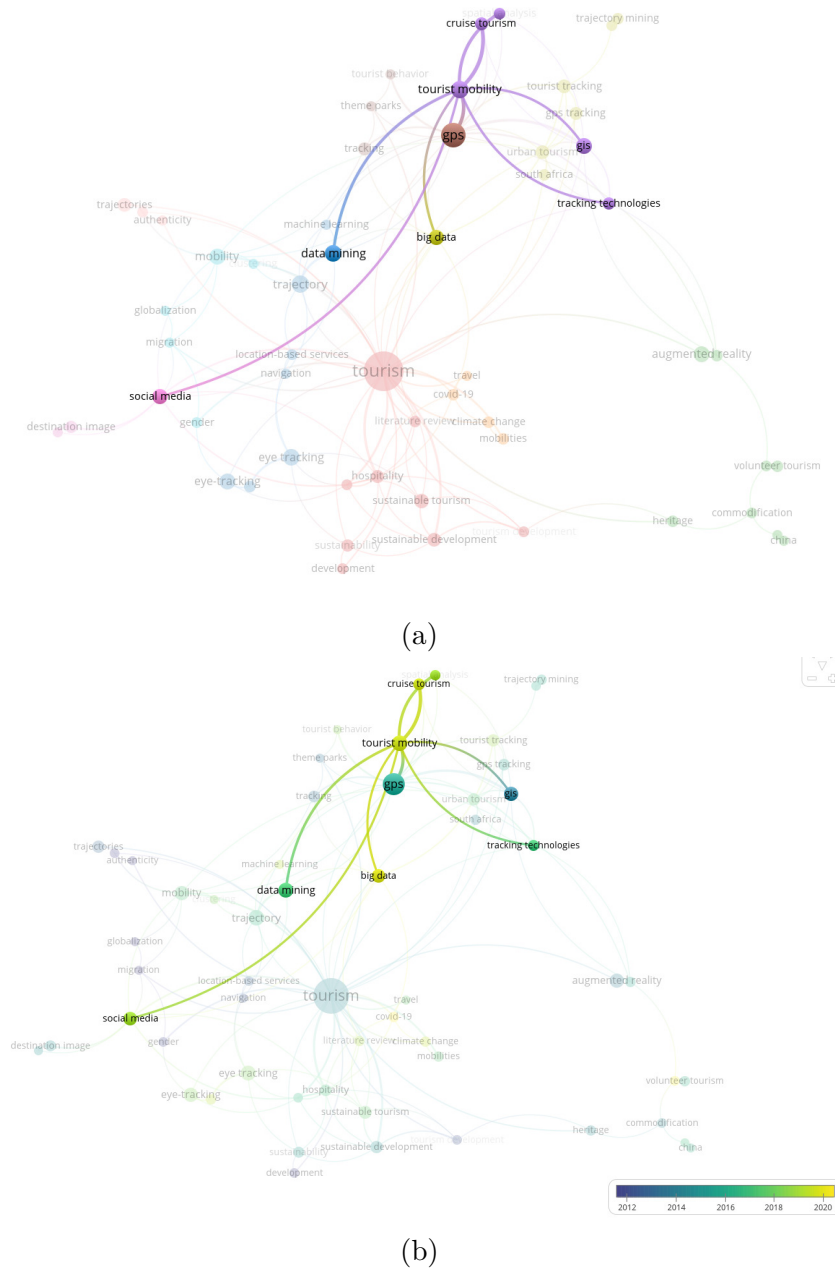


Fig. 4.3 The cluster of analysis of tourist mobility with: a) GPS and tracking technologies; b) trends.

These three clusters concluded that the tourist tracking models were based on the domain context, such as urban environments, cruise ships, or theme parks. Technologies such as GPS and GIS are used to carry out the tracking. Social media sources are also used to obtain data for tracking, and in this way, analytics from technologies such as Big Data, data mining, and spatial analysis was applied. Finally, the applications of tourist tracks

can vary from the analysis of the flow of tourists, trajectories to the tourist’s behavior on their touring.

Table 4.1 lists some literature review studies that were obtained from systematic mapping. The technologies and tools considered for tracking the tourist for each study were identified. The subject column describes the different issues involved. The paper [121], although its domain was not tourism, was taken into account due to the in-depth analysis of the theory on pedestrian trajectory prediction, which is closely related to the purpose of this research.

Table 4.1 Some studies of literature review about tourists tracking

Research	Technologies and tools	Subject
Wang et al. [122]	Sensor, mobile, GPS, RFID, Surveillance cameras, and, UAV	Trajectories and trajectory data management system
Sighencea et al. [121]	Advanced Driver Assist System (ADAS), DL, DNN, radar, LiDAR, video camera, Self Driver Vehicle (SDV), and sensors	Pedestrian, trajectory prediction (PTP) methods
Padrón et al. [124]	Survey, Web analysis, geolocation, advertising, sales, and specific spots	Tourist tracking techniques
Shoval and Ahas [74]	GPS, mobile, Bluetooth, social media, and photos	Tracking technologies in tourism
Thimm and Seepold [125]	GPS logger, smartphone, app, and survey	Tourism movement patterns via tourist tracking

Table 4.2 Some studies of tourists tracking and trajectory approach

Research	Subject	Approach	Technology
Zheng et al. [126]	Tourist trajectory data analytics	Adaptive spatial clustering and Frequent pattern mining	Mobile
Park et al. [127]	The similarity of travel trajectories	Data mining and graph-based spatiotemporal analytics	Cell phone towers and roaming
Chu and Chou [128]	Tourist trajectory analysis	Region design model, spatial tourist networks, betweenness centrality (BC), brokerage analysis	Social Networks Analysis (SNA), Mobile phones, Call Detail Record (CDR), and GIS
Chen et al. [129]	Mobility tracking	Big Data and spatial-temporal resolution	Weibo
Mikhailov et al. [130]	Car Tourist Trajectory Prediction	Bidirectional LSTM Neural Network	Smartphone and IoT with ambient intelligence technologies
Park et al. [131]	Spatial structures of tourism destination	Trajectory data mining, spatial clustering analysis, and sequential pattern mining	Mobile and Big Data
Cayère et al. [132]	Tools for processing digital trajectories of tourist	Processing and analysis of digital spatiotemporal. Extract, Transform, Load (ETL)	Mobile application
Eccleston et al. [133]	Tracking technology for tourism planning and development	Data analytics, survey, dashboards, and reports	App and GNSS
Buning et al. [134]	Tracking visitor spatiotemporal behavior	Big Data	GPS
Ferrante et al. [62]	Framework for collecting and analyzing the tracking data	Surveys	GPS
Sakouhi et al. [135]	A Mobility Data Model for Web-Based Tourists Tracking	Tourist Mobility Data Model (TMDM), sub-models: Touristic Data Space (TDS), Event Space (ES), and Tourist Space (TS)	Websites users data and DATA-tourisme ontology
Chen et al. [136]	Mobility prediction based on POI-clustered data	Term Frequency – Inverse Document Frequency (TF-IDF) Processing, K-Means Clustering, and Dirichlet Process Mixture Model	Cellular network data
Jang et al. [137]	Navigation Tracking Systems	Beacon Detection and Direction (BDD) Indoor Positioning Systems (IPS)	Beacon
Shoval et al [138]	Implementation of tracking	Analytic	GPS and GIS
Tiwari and Kaushik [139]	Interesting locations (POIs)	Fuzzy inference system (FIS) and ratings	GPS, GeoLife dataset, and HERE maps

These studies have taken advantage of existing technologies to track tourists in keeping with the epoch. Its main objective was to analyze the flow of tourists within a destination [129, 132, 131, 136, 139]. GPS was the predominant technology in these studies, which dealt with issues such as location accuracy, which was improving over the years, from the first GPS loggers accompanied with GNSS to the precise microsensors present in current smartphones and specialized positioning devices. With the primary data: geocoordinates based on latitude, longitude, and altitude, and with the time stamp, systems supported by the trajectory theory could be established. Nevertheless, the imprecision of the first GPS devices, coupled with the intermittence of GNSS, made researchers opt for more traditional tools to trace trajectories, such as surveys, travel diaries, or related systems [136, 138, 139, 62, 134].

Thanks to the massification of cellular mobile telephone systems, the tracing of tourist trajectories was positively affected by CDRs and GIS. The location of the CDRs is more precise by the cell systems that characterize these telephone systems [126–128]. Thus arose studies, especially by Asian researchers, who had vast amounts of data from telephone operators at their disposal. Therefore, the applications of these investigations were more varied to the analysis of the flow of tourists. The investigative interest in the behavior of the tourist in his touring appeared. In other latitudes of the planet, this trajectory data was limited, or even canceled, by the provisions on the processing of personal data [135]. Therefore, the new research challenge arises of opting for other data sources, even by combining them (cross-device tracking) to continue tourist tracking studies [128–130, 137].

Investigations have given a subsequent advance to the flow of tourists and their behavior in the destination. We are now interested in the bidirectional relationship of the tourist with the POIs [128, 133]. Since this is the consequent interest of our study, the tripartite relationship between tourist tracking with the destination and, in turn, besides the tourism value chain: a tourist traceability system comprises these three components. The first component, tourist tracking, comprises a spatial relationship, the second component, the trajectory of the tourist, establishing spatio-temporal causality, and the third component, the relationship of historical behavior and feedback. Therefore, we considered that tourist traceability is the natural evolution of tourist tracking and tourist trajectory analysis systems.

4.2 First Tourist Traceability Alternative

The first tourist traceability alternative involves those studies that considered the analysis of the destination and its attributes. For example, [136] collected data from POIs and applied

processing methods to obtain their socio-economic activities and functional attributes. To do this, they used the TF-IDF statistical method, accentuating the meaning of the words to categorize the documents through the vectorization of areas, which represented the levels of importance of their functions. Afterward, the TF-IDF vectors of the area units were clustered according to the type of functionalities, using grouping methods, for example, K-Means.

Once the destination profile has been considered with the methods described above, they proceeded to determine (or predict) the trajectories of the tourist, that is, to try to establish a relationship of the tourist's spatiality. For this, the behavior patterns of the tourist moving from one point to another are considered, forming a trajectory. These movement transitions within a trajectory can be represented mathematically in different ways, for instance, with Markov chains of order-1, where the trajectories are drawn with standard kernels to determine the probability that a tourist moved from a *point_i* a *point_j*. If the amount of data is insufficient, temporal patterns of users are reassembled, supported by Bayesian mobility models. Finally, using cluster-based mobility predictor algorithms, predictions are made using the trajectory history in a specific range. Thus, the visualization system is based on the representation of clusters (See Figure 4.4).

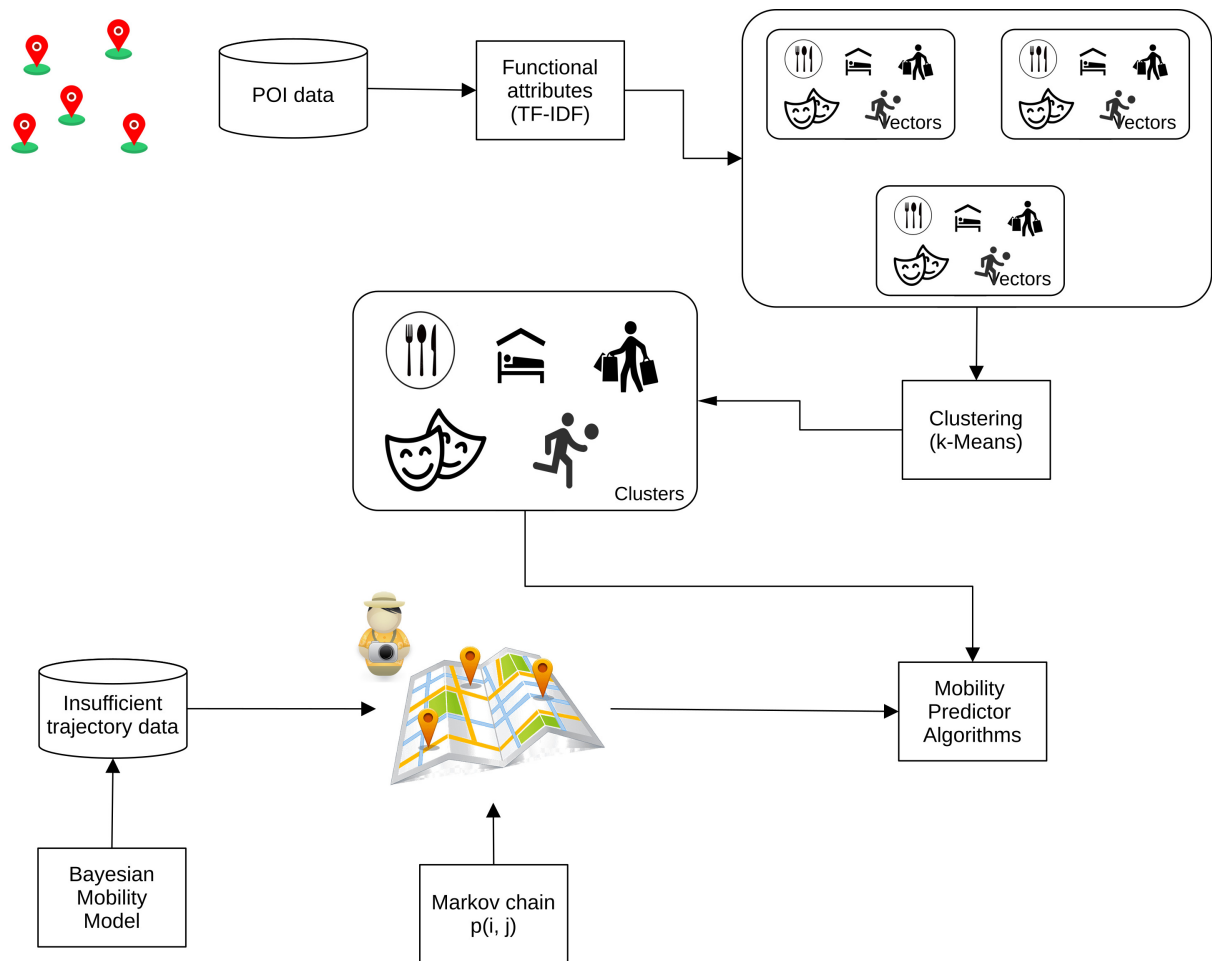


Fig. 4.4 First Tourist Traceability Alternative

4.3 Second Tourist Traceability Alternative

The second alternative includes Big Data and data mining to deal with as much data as possible, especially data from sources such as social networks, sensors, and trajectory data from mobile phone providers. Therefore, the method was based on stages similar to Big Data analytics. In the case of [127], a data collection stage begins, where the data sources, the data structure, and its cleanliness are analyzed. Subsequently, in the data analysis phase, the study of trajectories was incorporated through their stages, such as synthesis, matching, sequence of similarity, and estimation. However, these methodological steps can be adapted, added, or eliminated depending on the research domain.

Finally, the studies devoted special attention to the stages of analysis of results and data visualization, where the items to be reported, their relationship, and impacts are defined. The identification of movement patterns and trajectories to predict the tourist's flow within

each of the POIs in the destination is also highlighted. Apart from decision-making, the researchers agreed that the applicability of these systems is enormous, but their implications must also be had in mind (See Figure 4.5).

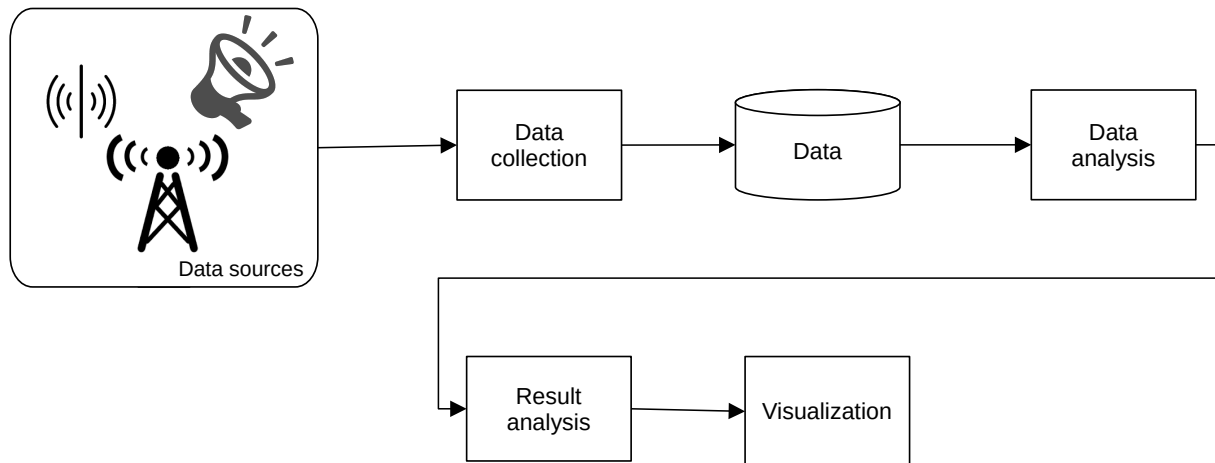


Fig. 4.5 Second Tourist Traceability Alternative

Chapter 5

Tourist traceability system model

Sections 2 and 3 provide an overview of a TTS. For proper decision-making by DMO, the TTS have components specialized in generating adequate information. This section proposes a model as a new alternative to TTS (see Figure 5.1). This model comprises subsystems that allow data processing from ubiquitous sources, a knowledge base called OntoTouTra, and subsystems that visualize and report decision-making information.

According to the tourism value chain analysis, the data collection and pre-processing stages are of utmost relevance for the TTS operation. Once the data had been obtained, we determined the tourists tracking, possible trajectories, and bidirectional relationship with the services offered. The location intelligence subsystem processes location data to determine the tourist's tracking (latitude, longitude, and timestamp). Consequently, we defined the opinion mining subsystem, which extracts valuable data through NLP techniques and detects hidden patterns of reviews. This system requires a central knowledge base that allows the TTS information to be structured and retrieved. With the data obtained from the two previous subsystems and the ontological design by the planning and designing stakeholders of the destination, OntoTouTra was created. This ontology is the backbone of the proposed model because all the subsystems place and obtain information and knowledge from the TTS. Following subsystems were the visualization, the dashboard, the decision-making, and the generator of the portfolio of tourist experiences of the destination.

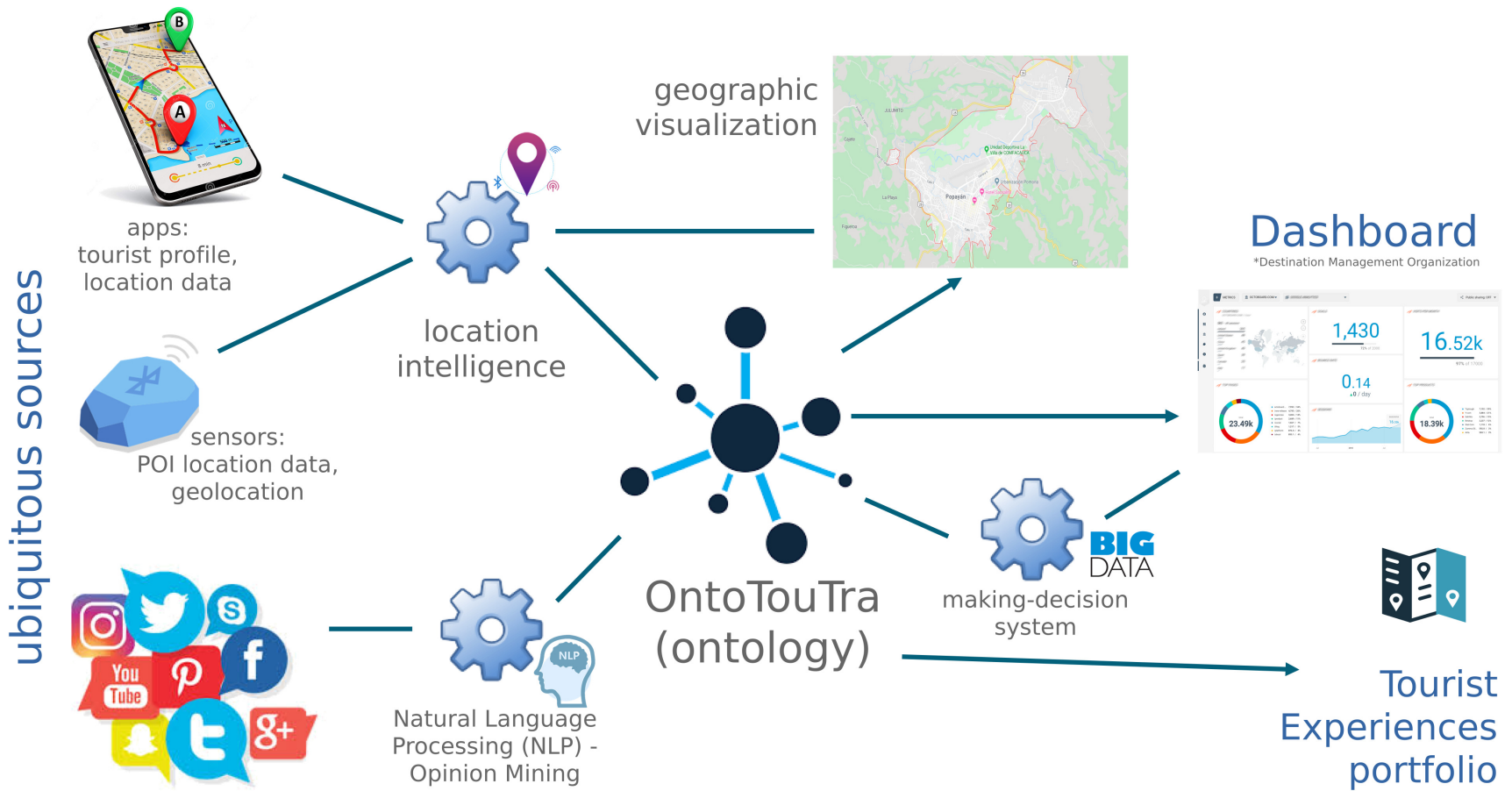


Fig. 5.1 TTS model

5.1 Ubiquitous data sources

Ubiquitous computing is the cross-device mechanism from which we obtain different data types. Our model (see Figure 5.1) starts from this challenge, that is, from what sources and how a large amount of data is obtained, in multiple formats, at different velocities, and with different levels of value. We used data collection techniques based on Big Data to solve this challenge. In a tourist environment, multiple ubiquitous sources can be identified, but initially, in our model, we considered three sources, namely:

- **Social networks:** The OTA strategy is the "push and pulls" that have created a symbiotic relationship between the tourist and the company. "Push" because the OTA encourages tourists to use its system, for example, booking, tourist plans, offers, tourist guides, news, among others. While in the "pull," the OTA motivates to get the tourist feedback, especially reviews, ratings, and even user networks. In the tourist's reviews, we found a data source with innumerable possibilities, but its gathering, pre-processing, and treatment require a higher level of complexity within the TTS model. The challenge arises of collecting the data from these devices, for which the IoT provides different hardware and software strategies, such as the configuration of gateways, buffer memories, transmission by bursts to the Internet, or different types of connections to mobile devices.
- **Sensors installed in the POI:** Data supplied can be straightforward (latitude, longitude, altitude, and timestamp), processing can be advantageous, especially for calculating trajectories, as explained in Section 4. The context settings of these sensors are different, such as edge computing, cloud computing, and personal networks (PAN).
- **Tourist mobile device data:** When using tourists' mobile devices as the third ubiquitous source, it is imperative to have prior authorization to use and treat the data collected from these devices. Likewise, make the app user aware of the purpose of data collection and guarantee confidentiality, integrity, and availability (CIA triad). In this way, from mobile devices as a bridge to other ubiquitous sources, we can obtain data on the context, the tracking, the trajectory, the profile, interests, and all those data of the tourist's relationship with the services offered in the value chain of tourism. The tourist requires added value by allowing the installation of this type of apps on their device, so it is necessary to plan a value strategy to awaken the interest of tourists for these applications. Within these strategies, there are discounts to enter the attractions, priority in attention, contextual information presented on the device (supported by technologies such as augmented reality), and access to multimedia

resources. Finally, this ubiquitous data source is noteworthy, since the technological advance of these devices is dizzying, the sensors that include (GPS, accelerometer, barometer, gyroscope, biometrics, pedometer, magnetometer, proximity, capacitive sensors, among others) generate data of great importance that can be correlated with other ubiquitous data.

5.1.1 Social network data

The OTA selected for the design and testing of the TTS was Booking.com. This OTA has maintained its leadership among its competitors due to the solidity of its platform, the number of registration of accommodations and users, and the alliances it has formed with different providers of the tourism sector worldwide (See Section 5.4.11, Table 5.4). To collect the data from the OTA, we created a Web Scraping application in Python that works through filters, especially one of the regions to be scraped, for example, a country or a region. It starts by scraping the destination data, the accommodation data, and its services. In a subordinate way, it collects the data underlying this region. Finally, we got the review and rating data. The data format depends on the unit of information retrieved: structured, semi-structured, and unstructured. Reviews are an example of this latest data typing. Due to its variety, volume, and velocity of obtaining, we store the data in a NoSQL database (MongoDB). The main class structure of the application can be seen in Figure E.2. The data obtained through WebScraping were essential for the design of the ontology and the generation of individuals for it (see Section 5.4.12).

5.1.2 Sensors

A wide range of sensors can be installed at destination attractions to capture location and context data. In this study, we installed a cluster of beacons in a POI. Beacons are hardware transmitters that broadcast their identifier to nearby devices. Beacons technology allows devices to execute actions depending on their proximity to the beacon. We used Bluetooth Low Energy (BLE) beacons, a personal wireless area network technology characterized by its low power consumption and because it is a low-cost technology. BLE uses the 2.4 GHz frequency, the same frequency as classic Bluetooth. In this frequency band, the devices are configured in dual-mode to share the same radio antenna through a simple modulation system.

An app (Electronic leash, always-on) is installed on the user's mobile devices. In our case, we developed two versions of an app called MEB, the first with Phone Gap and the second with Android Studio, in the Kotlin programming language. We configure the two

types of profiles: FMP (Find me profile) that allows you to be alert, and PXP (Proximity profile) that monitors the proximity of the other device in a close range. Although not exact, the proximity can be estimated from the radio reception value. For dealing with this inaccuracy, it is necessary to calibrate the absolute distances using a beacon map.

The transmission protocols of the beacons that we installed were iBeacon (Apple), and Eddystone (Google). With the beacon network configured, a positioning system was created within the POI (Indoor positioning system) to determine the approximate location of the smartphone within the context.

For the indoor location of the beacons, first, we mapped the location (See Figures B.1, B.2, B.3, and B.4). We defined the shape and Figure using a metric system (EILPoints), then we located the beacons on the site. Subsequently, we calibrated the beacons using metrics and orientation. The manufacturer of the beacons used has its platform (Estimote cloud). Despite this issue, we collected the proximity data through the developed app, which through an API, the App ID, and the token values were configured in the application (See Figure B.5).

For the test of network beacons, we developed the application for smartphones. Initially, the project was created, the app was connected to the cloud, the location was fetched from the cloud, the Indoor location manager was built, and finally, the position updates were started. The source code was adapted from the code supplied by the manufacturer (see Figure B.5). The execution of the application can be seen in Figure B.6.

5.1.3 Mobile app

An experiment was created with 18 participants who made their typical tours within the city, especially the POI chosen for the experiment. This POI was chosen for its cultural and heritage value in the city. An app developed by us, called MEB, was installed on each participant's smartphone. The app allowed recording the emotion and the activity that the participant felt and carried out at a particular moment. When the participant chose the emotion and the activity internally, the app recorded this data along with a timestamp, latitude, and longitude. This data was obtained from the smartphone's sensors. These data were collected for eleven weeks. Then, the data was synced with the Firebase cloud platform. In turn, when the mobile device approached one of the beacons installed in the POI, it generated a new location record.

The application interface can be seen in Figure C.1, a sample of log storage in Figure C.2, and the number of records was 21,000.

5.2 Location intelligence

Location data is treated in a geographic dimension on the Earth’s surface. These data are commonly referred to as coordinates (latitude, longitude, and altitude). Therefore, there are some variants of nomination such as postal zones, addresses, regions, cities, among others; the location intelligence subsystem processes geographic (spatial) data for obtaining insights.

This subsystem has components for obtaining geographic position data, location-sensitive preprocessing, analytical methods, and visualization. Through this subsystem, we can find hidden patterns in spatial dimensions.

The TTS model handles spatial data by vectors, represented by points, lines, or polygons. However, the model stores raster information through cells that handle continuous data, such as satellite images and aerial photographs.

Establishing the relationship between two points to determine their distance is known as the Great Circle Distance (GCD), which is based on the Haversine distance, similar to the Euclidean distance that establishes the minimum distance between two coordinates, with the exception that it is in account of the spherical nature of the Earth. Figure D.1 shows the Haversine distance calculation implemented in Python.

However, in urban environments whose planning is based on blocks, limited by street crossings, algorithms based on the distance from Manhattan are often used. In figure D.2, we see an implementation of the Manhattan distance with rotation. This distance is calculated as the sum of the distance in a straight line along the x-axis and the straight-line distance along the y-axis.

In the tourist domain, depending on the context (blocks or cross country), we can apply one of the two algorithms or a combination of both to calculate the distances of the trajectories.

With the data collected from the ubiquitous sources, we extracted the location data from the NoSQL database and formed a dataset (See Figure D.3), and through Map-Reduce operations, we executed operations within the Location Intelligence subsystem in Big Data environments. Figure D.4 depicts a code snippet in Python.

Tracking an object on a spatial plane generates a series of ordered points representing a path. Therefore, a trajectory is composed of two or more spatial points, defined by the Equations 5.1 and 5.2 [122].

$$p = (x, y, t) \tag{5.1}$$

where:

p = point;
 x = latitude;
 y = longitude;
 t = timestamp;

$$T = (p_1, p_2, \dots, p_n) \quad (5.2)$$

The number of points determines the length of the path, and the sample rate is the number of samples per unit time. Operations such as Cleaning, Storage, Similarity with other trajectories, Indexing, query, and analytics can be applied to the trajectories [122].

5.3 Opinion mining

Opinion mining applied to tourist reviews allowed to identify several essential aspects in a TTS, for example, tourist satisfaction by determining the polarity of the reviews, the object of the review (destination, POI, provider, among others), the service used, the tourist experience acquired, the recommendation to other tourists, the profile of the reviewer (gender, country of origin, language), among others. The reviews are unstructured data, and we extracted these types of characteristics through NLP techniques.

In turn, we consolidated the ontological design and generated individuals (instances) of the ontology through opinion mining. In Section 5.4.13.1, the use of NLP for the design of OntoTouTra combined with Big Data Semantics techniques is explained in more detail.

5.4 OntoTouTra

5.4.1 Introduction

The relationship between the concept of traceability with the tourist contributes to the improvement of the methodological approaches used in the studies because it provides us with the precision and validity of the data obtained, especially from ubiquitous environments [103]. Traceability constitutes an advance for the collection of tourist mobility data in spatial-temporal relationships. Traditionally, in the fields of production, logistics, or software, traceability has been considered as the set of actions, metrics, and technical procedures to identify and record each product from the beginning to the end of the supply chain [140]. Furthermore, the International Organization for Standardization (ISO) defines the traceability concept "as the ability to trace the history, application, or location of that which is under consideration." [51]. Also, the GS1 defines tracing as "the ability to identify

the origin, attributes, or history of a particular traceable item" and tracking as "the ability to follow the path of a traceable item." [141].

In this sense, through a TTS, the DMO can identify the routes of the tourists and the degree of interest that the attractions of the destination arouse in them. Also, TTS can use socio-demographic metrics and statistics reports to identify tourist profiles, prepare and adapt both the tourist destination and the tourism management system. Hence, with the accelerated technological advance that characterizes ubiquitous computing, now DMO have at their disposal various data sources. These sources provide input data for TTS, such as social networks, cloud platforms, Web, Internet of Things (IoT), traditional databases, public or private datasets, and linked data, among other data sources.

On the other hand, these data sources typically are extensive volume data sets and reach high speed (in real-time or almost in real-time). Also, variety is another characteristic of these data (some have a format, the vast majority do not). Big Data can process and store this type of data and constitute a knowledge base through ontological systems. In this way, DMO can make decisions based on the information processed.

Currently, in most cases, DMO make decisions based on paper surveys applied to some tourists. Also, government reports and those of the tourism sector actors serve as data for this decision-making process. These strategies have drawbacks, such as the subjectivity and predisposition of tourists to answer surveys. Many of them prefer not to answer them for time or data privacy reasons, government reports are generated in extended periods, and in some cases, they arrive late. For this reason, the research gap of this study arises, which takes advantage of data from ubiquitous sources to provide information related to the traceability of tourists to a given destination. In this way, with the processing of these characteristic Big Data data, precisely due to its volume, velocity, and variety, to constitute a knowledge base, the research question arises: How to develop a tourist traceability ontology based on obtaining and ubiquitous data processing, using Big Data analytics techniques?

It is worth mentioning that the purpose of this study is to constitute an ontology-based on data previously generated in a massive way, not on data from tourists in particular. Initially, we consider the data from three types of ubiquitous sources: Reviews of tourists in OTA, data from sensors located in the POI of the destination, and data from tourist guide applications installed on the tourist's mobile devices, which have prior permission for further processing. A tourist traceability ontology allows DMO to make decisions regarding the management of the destination according to the flow and track of tourists, determine their preferred POI, intelligently dispose of the infrastructure for adequate attention, foresee improvements in services. Furthermore, design tourist experiences according to the interests of the tourist in a space-time causality.

The **OntoTouTra** is an ontology that explains the structure of knowledge, whose domain is the tourist traceability system, based on data collected from ubiquitous systems [142]. **OntoTouTra** shares this knowledge through the conceptual design of this domain, enabling the reuse of knowledge.

5.4.2 Tourist Traceability System

A **TTS** can provide information to answer questions. Some of these questions are:

- **POI:** What are the busiest POIs? What type of visitors frequent them? In what time slot are they visited? Where do the tourists come from? Later, where do they go? What activities do they mostly do? What tourist experiences are enjoyed?
- **Seasonality:** What is the behavior of seasonality in the destination? What activities are carried out due to seasonality? What services do they consume due to seasonality? What is the offer of tourist experiences?
- **Suppliers:** What is the level of satisfaction with the services provided? What are the needs to satisfy the demand?
- **Stakeholders:** How do stakeholders interact at the beginning, during, and at end of the visit to the tourist destination? What suggestions do tourists have regarding this service chain?

5.4.3 OntoTouTra Analysis

The domain of this ontology has as its main classes: the DMO, tourism experiences, tourist attractions (POIs), and destinations. We established the relations within the tourist domain:

- DMOs provide the service that the tourist consumes;
- The tourists live the experiences in the destination;
- The tourist attractions are the push factor and motivator for the tourist;
- The destination is the geographical location where tourist traceability happens.

These four relationships were the starting point of the ontology design; we designed the use case diagram (see Figure 5.2) and created the primary classes of the ontology mentioned

above. From these classes, we generated the subclasses, properties, and relationships between classes.

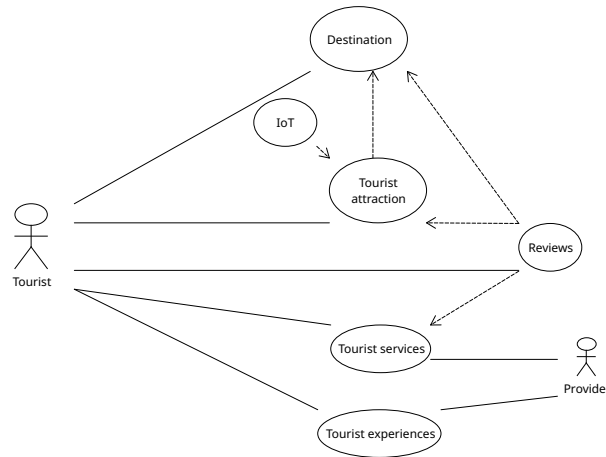


Fig. 5.2 Tourist traceability system: use case.

The top-down design of the ontology provided the hierarchical order of the terms, starting from the root domain, that is to say, the tourist traceability (root node), and distributed by the general classes until arriving at the specific terms. Identifying the terms from sources of authorities on tourism and other similar ontologies and with the DMO's expertise thus ensures the formulation and definitions of the ontology's taxonomic hierarchy. The process of the revision of iterative versions is necessary to guarantee the consistency of the definition and the scientific, logical, and philosophical rigor of the terms (see Figure E.1).

5.4.4 Development of the Ontology on the Domain of TTS

We focused on using a method for the ontology's construction, such as METHONTOLOGY [143], a methodology that allows building ontologies from scratch and has been tested in different knowledge domains. Using this methodology, we took advantage of the Big Data analytics lifecycle model [4] to obtain, process, classify, and visualize data from ubiquitous computing sources. This ontology, called OntoTouTra, has as its principal purpose to provide a knowledge base to handle problems of semantic aspects to support the implementation of a TTS.

METHONTOLOGY [143] uses an iterative approach to tailor the ontology to refine the TTS domain model. In this way, we moved from the level of knowledge (conceptual model) to the level of implementation (logical or computational model), looking for the ontology to be readable by machines. For the construction of the conceptual model (see

Section 5.4.10), we began with the identification of the purpose and scope of the ontology, as follows:

5.4.5 Specification

The domain of the ontology is the TTS with four main branches: Provider, Tourist Experience, Destination, and Tourist (see Figures E.1 and 5.3). Understanding these branches avoided any inconsistencies between the classes and the ontology. In addition, these branches responded to the requirements of tourism traceability: Where are the tourists (Destination)? What do tourists do (Tourist Experiences)? Who offers the experiences (Provider)?. From these branches were derived the classes that make up the TTS domain. The POIs and the tourist reviews are important because they implemented the space–time relationship to answer questions of the domain: When and where does the tourist consume the experience, or what is the tourist’s opinion of the experience?

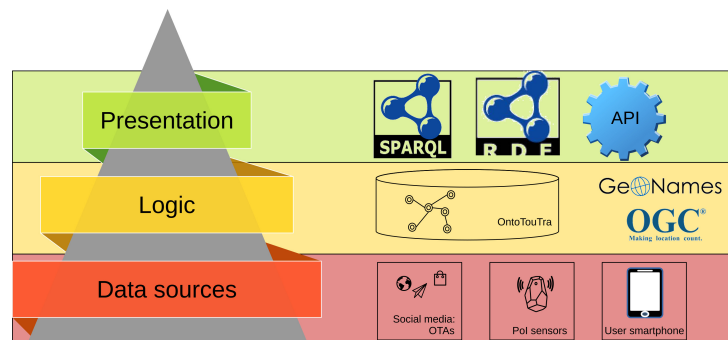


Fig. 5.3 OntoTouTra architecture

5.4.6 Conceptualization

We considered two types of data sources for knowledge acquisition (see Table 5.1). The first source corresponds to expert organizations in the tourism domain. We analyzed the management documentation, policies, guidelines, and reports to define the ontology domain’s branches, classes, subclasses, relationships, properties, and scope. The second data source is ubiquitous computing. In this phase, we refined the ontology, iterating between the specification and the conceptualization. The data were gathered with web scraping from the OTA. Due to its Big Data characteristics, we applied the appropriate methods for these environments, such as data mining, text mining, and the MapReduce technology. With the vocabulary obtained (see Section 5.4.13.3), mainly from tourist reviews, we refined the previously defined concepts with the first data source type.

Subsequently, we built the TTS glossary, identifying the concepts and ensuring each term was described with synonyms and acronyms. We also checked the terms that referred to the same concept (related terms). Each term has a simple description within the ontology. Through the relationships between classes, we avoided any ambiguity of concepts, for instance: destination-city-municipality, point of interest-POI-attraction, tourist-visitor-reviewer, and provider-supplier. A fragment of the TTS glossary is depicted in Table 5.2.

We implemented the top-down approach (see Section 5.4.10), starting with a general level until the level of the details. By identifying the classes and their relationships, we defined the taxonomy and hierarchy of the ontology. According to Kumara et al. [144], a *hierarchy* is defined as $H = (N, E)$, which is a simple directed graph, where N is the nodes and a set of edges $(n_p, n_c) \in E \subseteq NxN$. The address of an edge (n_p, n_c) is defined from the parent node n_p to the child node n_c (SubClass Of).

Another type of relationship between classes describes their behavior. For instance, the class “Hotel” has a relationship “*hasService*” with “Service.” For example, from the class “Provider,” we obtained several subclasses, according to the category of the service offered, so the class “Hotel” has an include relation of “SubClass Of” from the class “Accommodation”, and this, in turn, has a relation “SubClass Of” with “Provider.” This last relationship is an illustration of ontology refinement using data-mining techniques in Big Data environments.

Table 5.1 Data sources of the individuals of the main classes of OntoTouTra.

Ontology Main Class	Data Source (Individuals)	Linked Data	Data Sources Used in This Research
Tourist	social networks: OTA, eWOM	foaf	[145], [146], [147], [148]
Experience	tourist providers’ datasets (DMOs)		MinCIT-Open Data [149], DataEco [150]
Provider	government providers’ datasets		MinCIT [151]
City	social networks	GeoNames	[145]
Attraction	social networks, IoT (POI wireless transmitters)	GeoNames	[145], beacons
Hotel	social networks: eWOMs, OTAs		[145]
Review	social networks: eWOMs, OTAs	time	[145]

Table 5.2 Glossary of a TTS (sample concepts).

Term	Synonym	Acronym	Description	Type
Attraction	Point-of-Interest	PoI	A place of interest where tourist visit for its value or significance.	Class
Tourist	Visitor		A person who travels away from their normal residential region for a temporary period of at least one night, to the extent that their behavior involves a search for leisure experiences from interactions with features or characteristics of places he/she chooses to visit.	Class
Tourist experience		TE	A set of activities in which individuals engage on their personal terms, such as pleasant and memorable places, allowing each tourist to build his or her own travel experiences so that these satisfy a wide range of personal needs.	Class
Destination	City		A geographical area consisting of all the services and infrastructure necessary for the stay of a specific tourist or tourism segment.	Class
Provider	Supplier		All businesses offering tourism services and experiences to consumers when the latter are traveling and performing tourism activities.	Class
Review	Opinion		A subjective opinion of a tourist's experience.	Subclass

5.4.7 Formalization and implementation

We used Protégé as the editor and framework for the construction of OntoTouTra. Through formalization, we produced meaningful models at the level of knowledge. We gave each class or subclass term a semantic relationship between them (see Table 5.3). In this phase, we solved the semantic problems detected, for instance, the need to specialize the tourist experiences in subclasses and determine subclasses for the tourist reviews according to the provider or the geographic location of the review. The formal language used was OWL/RDF.

Table 5.3 OntoTouTra relationships (owl:topObjectProperty).

No	Relationship	No.	Relationship
1	belongs	9	hasService
2	enjoys	10	hasServiceCategory
3	hasAccommodationType	11	hasStateParent
4	hasCityParent	12	located
5	hasCountryParent	13	offered
6	hasHotel	14	operates
7	hasHotelScore	15	uses
8	hasScoreCategory	16	visits

5.4.8 Evaluation

At this stage, we verified the level of consistency and acceptance of the ontology knowledge. We did this process from three approaches. The first consisted of verifying whether the defined objectives met the purpose of the ontology. For this, we followed the FOCA methodology. The second was the validation of the conceptual model to determine the effectiveness of the ontology. To do this, we used the Competency Questions (CQ) approach by calculating ten KPI from a TTS system. The last approach corresponds to the test of the ontology through a use case. We generated the ontology individuals from web scraping of an OTA for the Colombian tourist case. We created ten test case scenarios with this case study and executed SPARQL Protocol and RDF Query Language (SPARQL) for each KPI from the previous approach. The results of these SPARQL queries were contrasted with the expected results obtained from the sources of authorities in tourism. Section 6.1 details each of these three OntoTouTra evaluation approaches. Besides, we made a document (see Appendix A) with the implementations and results of these test cases.

5.4.9 Documentation

The documentation is essential to recognize the current state and maintain the ontology's consistency. For this process, we used two tools for the automatic production of the documentation: Protégé and Ontology-based APIs (OBA) [152].

Regarding the logic model, the OntoTouTra architecture is multilayered based on functionality (see Figure 5.3, Section 5.4.10). As mentioned above, this architecture operates in Big Data environments, wherein the lower layers use data-mining techniques to process data from ubiquitous data sources. In the upper layer, the ontology offers different

data recovery possibilities, such as the traditional [SPARQL](#) queries from an endpoint and REST API requests, the implementation of which can be seen in the screenshots of the [Appendix A](#). Taking advantage of these ontology query possibilities, we handled scripts in programming languages, especially Python, to perform complex queries with Big Data analytics techniques, using the PySpark and PyMongo libraries.

5.4.10 Model for the Development of OntoTouTra

Our model for developing the ontology of tourist traceability has the following components (see [Figure 5.4](#)).

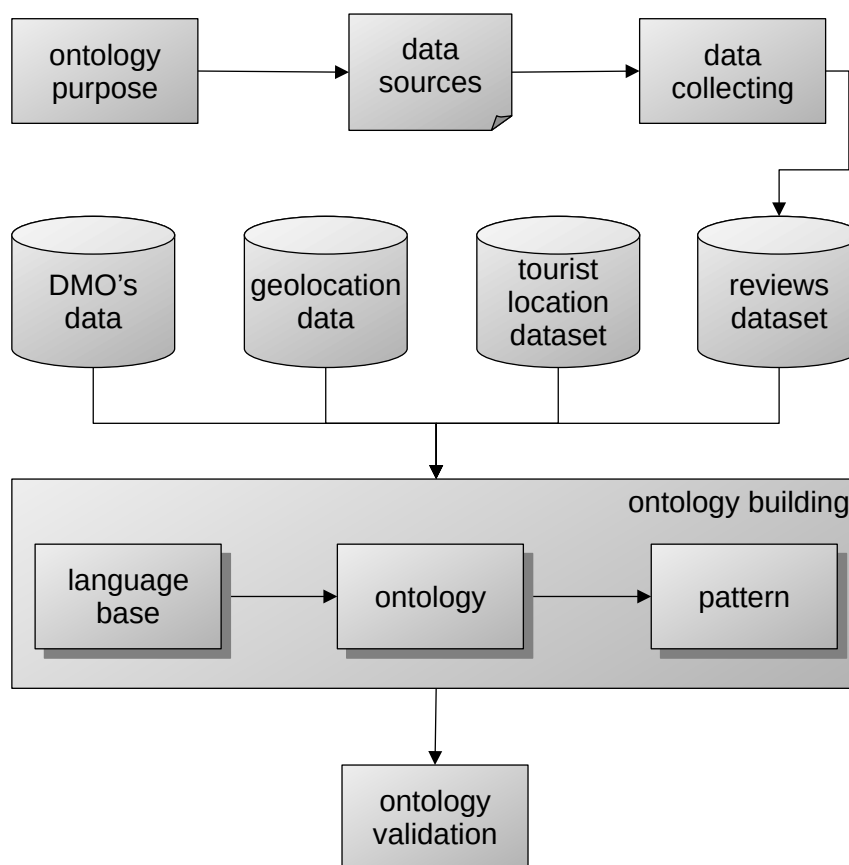


Fig. 5.4 OntoTouTra development model.

Next, we define and explain the procedure for each of the stages of the model that we developed to create and validate the OntoTouTra ontology through lists, diagrams, tables, and statistical graphics. We also provide the necessary recommendations to satisfy the requirements of each stage:

5.4.10.1 Definition of the ontology's purpose

The model begins with the scope of the ontology of tourist traceability, the justification, the motivation, and the goals. The purpose may arise from the need for decision-making by the DMO to improve the destination and its POIs. This component is mandatory.

5.4.10.2 Data sources

The sources from which the data are collected can be governmental, public, or private sources such as the regulations for the provision of tourist services, information systems, social networks, other ubiquitous sources, reports from the [UNWTO](#), other tourism authorities, tourism reports from local and national governments, hotel occupancy data, restaurant management, and the entities that revolve around tourists (see [Table 5.1](#)).

5.4.10.3 Data collecting

We can collect data from the identified sources, which can be manual, semi-automatic, or automatic. The data can be on paper, files, datasets, ontologies, information systems, social networks, sensors, mobile devices, and the web, among others. We can use custom applications to obtain automatic or semi-automatic data, whether in batch or real-time processing (see [Figure E.2](#)). Some ad hoc developments may be required, mainly to obtain specific terms about tourism subjects that will be part of the corpus and the lexicon of the ontology, for example, to collect social networks data, [API](#), or web scraping, then applying data-mining techniques.

5.4.10.4 Tourist location dataset

Tourist traceability requires the tracking of their geographical location. The calculations of the geographic positions, the permanence in the POI, and the destination can be performed by utilizing coordinates or even by semantic analysis, which determines a specific location. Therefore, the classes of the ontology must have subclasses or attributes that facilitate the determination of geographic coordinates (see [Figure E.3](#) and [E.4](#)). Ontologies and external geographic datasets can form this component. The ontology must interpret the terms of locations, mainly as nouns or names.

As can be seen in the results of the query listing, there are no terms for latitude and longitude within the terms of the [OntoTouTra](#) ontology for cities. Using a data link to [GeoNames](#), these terms are obtained. To avoid ambiguities with the names of the locations, we attempted to obtain from the ubiquitous data source the most significant amount of data that characterized the location of that geographic entity, for instance the type of unit:

country, state or region, city, municipality, and neighborhood, among others; geographic coordinates (latitude and longitude) and direction. Furthermore, we established the relation of ontological classes, for example, a city has the relationship “hasStateParent,” a hotel has the relationship “hasCityParent,” and so on.

In tourism traceability, queries on the geographical issue are needed, which OntoTouTra alone would not solve. GeoNames is a specialized ontology and is ideal for complementing geographic data that OntoTouTra lacks., for instance, to perform population-related calculations, such as the rate of tourism companies for every number of inhabitants. OntoTouTra makes linked data with GeoNames and retrieves the data of the number of inhabitants of a specific geographic area.

5.4.10.5 Tourist reviews dataset

The tourist reviews provide items for the ontology; they are terms frequently used in tourist slang and the valuable channel of communication and feedback for the tourist ecosystem. Reviews can be obtained manually, such as surveys and suggestion boxes, or automatically extracted from tourism social networks depending on the data source.

Forming a dataset of tourist reviews has many advantages and serves as a corpus of the ontology. For example, through Natural Language Processing (NLP), we can obtain the polarization of the reviews. We can also establish the traceability relationship, that is spatiotemporal.

In Figure E.5, we see the distribution of the scores the tourists gave to the localities (cities) that they visited, through the process and visualization of the dataset of tourist reviews of Colombia, in English. In addition, through NLP, we can tokenize the tourist reviews, and in this way, the ontology terms are achieved through a filter. NLP also allows classifying the terms. This component can enrich it with unsupervised-machine-learning techniques to cluster the terms. Some terms of the previous component may be wrongly spelled, poorly categorized, or not relevant to the ontology developed. We used a simple filtering method to determine the frequency of valid terms accepted by the ontology. More filters can be applied to search the quality of the corpus of the ontology.

5.4.10.6 Ontology input data files

We entered individuals into the ontology manually or automatically. The current version of OntoTouTra was designed in Protégé [153], using the Cellfie plugin [154]. We uploaded the individuals’ spreadsheets. Cellfie creates the axioms of the ontology, using transformation rules, as seen in Figure E.6. Regarding the two remaining data sources, and as we mentioned earlier, the ontology design recommends using ubiquitous data with Big Data analytics

techniques. Class instances such as tourism experiences and provider data are often in these formats and can be loaded into the ontology. Whenever possible, we recommend reusing knowledge through open link data for geolocation data, which is very sensitive for a traceability system.

5.4.10.7 Ontology building

base language -> ontology -> pattern. This is closely related to the first three of the five proposed phases (Specification, Conceptualization, Formalization, Implementation, and Maintenance) of the METHONTOLOGY methodology [143, 155]. The OntoTouTra ontology architecture is multilayered (see Figure 5.3) based on functionality, from storage (low-tier) to interaction (top-tier).

- Layer 1 corresponds to the input data, mainly from ubiquitous computing sources, such as social networks, sensors located at the destination, and users' mobile devices. This process was carried out through a data analysis pipeline, where we applied qualitative and quantitative techniques when examining the data to provide valuable insight. Data analytics provides the means to examine the EDA and CDA findings. Using EDA, we explored the data to find patterns and relationships among different ontology elements. Furthermore, through CDA, we obtained conclusions to specific questions of the tourism domain, based mainly on the simple observation of the data.
- Layer 2 is the logical layer, achieved by reasoning from OWL/RDF storage. The reason is limited according to the domain and range restrictions defined in the ontology. Using this layer, we can explain the content, apply queries, and verify the integrity of the ontology.
- Layer 3 corresponds to the presentation; OntoTouTra allows data visualization with different SPARQL endpoints, APIs, and graph visualization tools.

5.4.10.8 Ontology validation

This ensures that the ontology fulfills its purpose (first component). Steiner and Albert [156] suggested the validation of the content, application, and structure. The ontology must work appropriately according to its approach with the criteria of consistency, completeness, and conciseness. The validation of the functional ontology was performed on a set of domain CQ tests. The tests were implemented as queries of the individuals of the ontology. These tests were confirmed with the reasoning system. acOntoTouTra uses the Protégé reasoning system, as is the case with Hermit Version 1.4.3.456 [157].

5.4.11 Development and Usage of OntoTouTra in Big Data Environments

Big Data is part of a strategic initiative to design and execute business technology solutions backed by the analysis and management of large volumes of data through technology [4]. Big Data is an ideal solution for analyzing, processing, and storing data from tourist traceability systems. We needed to combine multiple unrelated datasets, analyze the data provenance, process large amounts of unstructured data, and look for hidden data patterns in a time-sensitive way. The analysis allowed understanding the data, examining it employing scientific techniques and automated tools to discover hidden behaviors and patterns. From massive amounts of data, without processing or structuring, the relevant information was obtained. A methodology is needed to handle the different requirements to execute Big Data analytics.

Ubiquitous data sources, such as social networks and the IoT, require massive parallelism to obtain the vast volumes of data, the data distribution, high-speed networks, and data mining and analytics. A tourist traceability system depends on the processing of these data; we were interested in knowing the activity of the tourist within the destination and its relationship with the tourist actors. The reviews are an excellent example of this interaction since they provided us with that fundamental space–time causality for traceability. An alternative to analytics is graph analytics, which uses an abstraction called a graph model. This model connects large volumes of data from different sources and in various structures. Graph analytics gather structured and unstructured data by coupling them into entity relationships. We can infer, identify patterns of interest, and deduce through an iterative approach to discover knowledge through this analysis. The same as the ontology, the graph model is straightforward since it is based on entities (nodes) and edges (relationships) [158].

5.4.12 Big Data Analytics Lifecycle for Building the TTS Ontology

This methodology allows planning and organizing the tasks, activities, and resources for data management. As a methodology, this research adopted the lifecycle of data analytics [4], divided into nine states (see Figure 5.5).

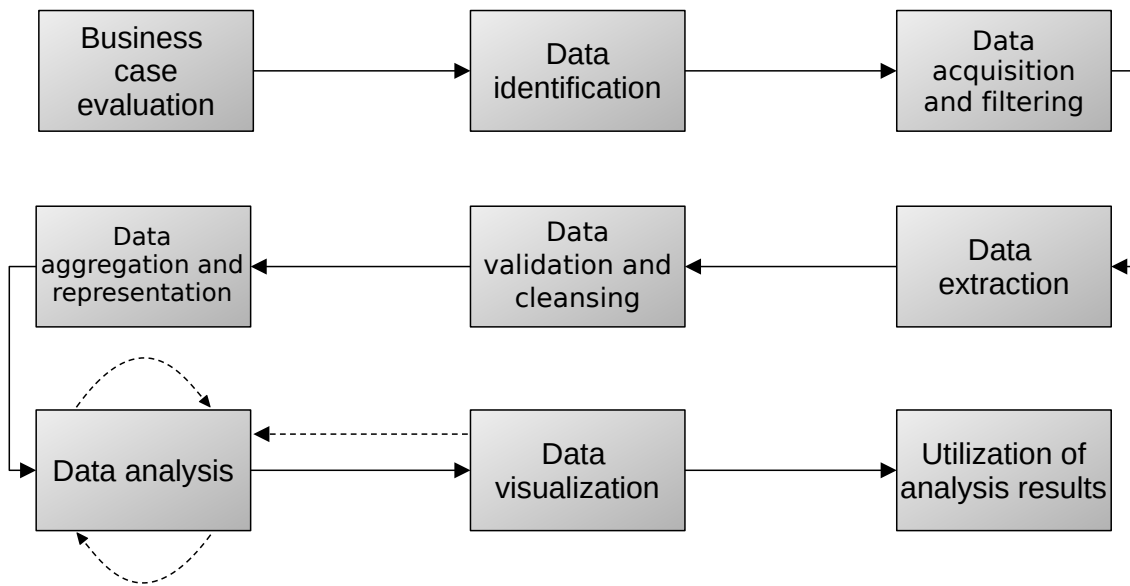


Fig. 5.5 Big Data lifecycle [4].

In 2016, Erl et al. proposed a lifecycle model for Big Data analytics [4]. It is a step-by-step methodology necessary to organize the activities involved in the acquisition, processing, analysis, and reuse of data. This methodology is applicable in any context. For this reason, we adapted these methodological phases for the construction and use of OntoTouTra in Big Data environments. Next, in each of the stages, we explain this adaptation, and employing some lists and charts, we demonstrate the implementation that we carried out in the ontology:

5.4.12.1 Business case evaluation

This is the stage related to the first and last component of the OntoTouTra model. It is necessary to have clarity about the justification, the motivation, and the objectives of the tourist traceability analysis. The motives to carry out this analysis can be various, among which we can mention: the marketing domain and the destination promotion, the actors involved in tourism management, the definition and application of policies and strategies, destination management, decision-making, and financial management [159] (see Figure 5.2). It is necessary to seek advice from expert Big Data and tourism management consultants because not all solutions meet the conditions and features of Big Data (the 5Vs: volume, variety, velocity, value, and veracity)

5.4.12.2 Data identification

In this stage, we determined the datasets and provenance. Location and tracking data for tourists within the destination are indispensable to satisfy the requirements of the previous step. Around the data, we required establishing the acquisition cost, confidentiality, and personal data treatment policies. Table 5.4 shows the leading OTAs that are potential data sources with information on the tourist domain. Booking.com registers the most significant number of accommodation listings for tourist site information and a tourist review platform. For many years, it has remained in the top 10 of the OTAs with the most excellent offer. In particular, in our case study, we chose this OTA

Table 5.4 OTAs (source: Cloudbeds, 2020).

OTA	Founded	Listings	Audience	Countries	Languages
Booking.com	1996	28 M	50 M	200	43
Skyscanner	2001	2 M	60 M	49	30
Expedia	1996	590 K	50 M	75	35
TripAdvisor	2000	7.3 M	490 M	48	28
Agoda	1998	2 M	2.3 M	65	38
Airbnb	2008	7 M	750 M	220	89
HostelWorld	1999	36 K	13 M	178	20
Hotelbeds	2001	180 K	60 K	185	20

5.4.12.3 Data acquisition and filtering

This involves gathering data by different means: files, digitalization, web scraping, integration with API, cloud services, transactional data, sensor data, information systems databases, and dataset providers, among others. Filtering is necessary to eliminate noise from the data. It is desirable to use data-mining techniques. In Figure E.7, we see the Python class that invokes the Selenium Web Scraping driver. The routes and parameters of the OTA were previously defined. Web scraping is hierarchically performed by region; for instance, we can start with a specific country and then go through its states or subregions. We also designed the class methods to obtain the information in a structured way from the hotels: general info, address, services, ratings, and reviews.

5.4.12.4 Data extraction

This extracts disparate data and transforms then into an understandable format for the Big Data solution. In the case of scanned documents, at this stage, it is determined if the

Big Data solution can read them in their original format or if we need to execute OCR applications. In the final part of Figure E.7, we see how the data are extracted and stored in memory in the datasets that were formed for each of the data structures of the hotels, tourist destinations, and tourist reviews;

5.4.12.5 Data validation and cleansing

Validation rules and removal of invalid data are applied to determine the accuracy and quality, for instance, the validation of destination geographic coordinates and the tourist activity timestamps. This stage is very demanding in a web scraping operation because the way the data are displayed on the OTA web pages can vary. Often, the information is missing, erroneous (due to user typing), or may be intermittent due to the conditions of Internet access to the site.

5.4.12.6 Data aggregation and representation

We required a unified data view by identifying the key fields to join sparse datasets because they come from different sources. This is a complex process because the syntax and semantics of the data model are determined. We designed this model with reuse principles for future requirements.

5.4.12.7 Data analysis

We set the ontology axioms and terms. Different types of analytics were applied to discover data patterns through operations such as queries, aggregations, or filters. The analysis can be confirmatory or exploratory, depending on the deductive or inductive approach. When we supply the ontology individuals (instances), exploratory analysis is the most suitable because it is closely related to data mining.

5.4.12.8 Data visualization

The results' interpretation leads to the formulation of the ontology, determining its structure (classes, relationships, functions, axioms or restrictions, instances, and properties or attributes), hierarchy, clarity, extensibility, and coherence. An example of the visualization of our ontology can be seen in Figures E.8 and E.9. A SPARQL geolocated query was executed on the ontology. Later, we stored it in a dataset, and using the `plotly.express` library, we visualized the results of the query on a map.

5.4.12.9 Utilization of analysis results

We built the OntoTouTra ontology, whose primary purpose is the knowledge base for a tourist traceability system. The results of the analysis can support decision-making for the tourism ecosystem. For example, we can apply machine learning and NLP techniques to determine the KPI of tourist satisfaction at the destination based on their reviews (see Figures E.10 and E.11). In Figure E.12, the polarity corresponds to the x-axis and the subjectivity to the y-axis. Polarity determines whether the review is positive or negative, while the size of the chart markers determines the subjectivity. We found more positive reviews located on the right-hand side.

5.4.13 Using Big Data

5.4.13.1 Components of the Analytics Toolkit

In this study, we utilized some key Big-Data-mining technologies to define the classes and terms of the ontology and build some queries. Table 5.5 shows the analytics toolkit used in this research.

The architecture diagram of the data pipeline can be seen in Figure 5.6. In the case of this diagram, we started by obtaining the data from a ubiquitous data source from an OTA (Booking.com), using web-scraping techniques with the Selenium library in Python. Then, we created the data flow of the respective data unit according to the scraping; we worked with the destination and its geographical coordinates, the data of the suppliers, especially the hotels, the tourist services, their ratings, and the tourist reviews with their temporality data. These streams were written as documents to a MongoDB collection. Subsequently, we built a Spark Streaming Dataframe that reads the MongoDB collection and periodically updates or adds new data. We made structured queries from the Spark Streaming Dataframe to store their results as axioms in the ontology. These axioms are of two types: The first type corresponds to detecting new patterns of data units that boost the ontology with new classes or terms (for example, a new attribute for the class “Provider” or a new class representing a tourist actor within the ontology). The detection of new data units for the ontology was carried out with NLP applied to the tourist reviews. The second type of axiom corresponds to the generation of new individuals in the ontology, such as, for instance, the creation of a new tourist experience, new groups of reviews, new hotel instances, or new tourist destinations or POI.

Table 5.5 Components of the analytics toolkit.

Software	Use	Function
Spark/PySpark	data mining	PySpark Dataframe for Big Data entities: reviews, hotel services, and scores.
MongoDB	data mining	Temporary storage for NoSQL collections, mainly tourist reviews.
Python	data mining/queries	Scripting for all functions: scraping, ontology API, loading of individuals, queries, and visualization.
RDFLib	queries	SPARQL API interface.
Selenium	data mining	OTA web scraping.
NLTK	data mining	Definition of ontology classes and terms. Analysis of tourist reviews for queries.

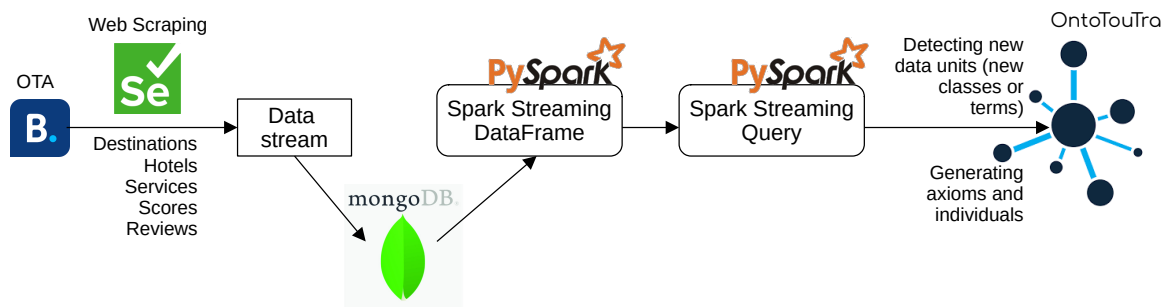


Fig. 5.6 Architecture diagram for the data pipeline.

To carry out complex queries that require the calculation of an enormous amount of data, we also used Big Data, for example, in the stages of data aggregation and representation, data analysis, data visualization, and the use of the analysis results, explained in Section 5.4.11 and Figures 5.5, E.8, and E.9; here, we executed the scripts on the *OntoTouTra* ontology and visualized the tourist destinations of Colombia on a georeferenced map.

5.4.13.2 Variety of Data

When applying web scraping, we collected data of various structures and, in some cases, without structure. In Figure E.13, we see the data flow of a tourist review in HTML code. We used MongoDB because it is a NoSQL database that stores unstructured data in the form of documents. In this way, using a Python script, we analyzed the data flow obtained from web scraping and converted it into JSON format for later loading into MongoDB, that is, we went from unstructured data to semistructured data. Subsequently, with the Streaming Dataframe and the Streaming Query, we generated the axioms of the individuals of the ontology in data in a structured way.

For the case study, the total number of *OntoTouTra* axioms for only one country depicts the Big Data volume feature, as shown in Table 5.6.

Table 5.6 OntoTouTra statistics.

Item	Count
Reviews	1,009,469
Services	481,443
Hotels	11,071
Destinations	678
OntoTouTra axioms	698
Logical axiom	352
Declaration axioms	190
Class count	65
Object property	16
Data property	109
SubClass Of	57
OntoTouTra axioms	17,225,580

5.4.13.3 Big Data Semantics

The relationship between Big Data and semantics is bidirectional [160]. On the one hand, Big Data's techniques and pipeline determine and filter the terms of an ontology and establish their relationships to provide the meaning of the domain. On the other hand, semantics [161] is a great tool to deal with the heterogeneity and variety of data. We can apply semantics in different phases of the Big Data lifecycle, such as detecting inconsistent data, discovering hidden patterns and data trends, and the data relationship necessary to create machine-learning models for different types of analytics: descriptive, diagnostic, predictive, and prescriptive. This bidirectional relationship manages large volumes of data at high velocity, and variety, thanks to the Big-Data-processing techniques. It provides meaningful, relevant, and valuable data for organizations due to the data semantics. The use of Big Data semantics in this research facilitated the generation of the OntoTouTra ontology in four aspects:

- The identification of relevant terms from a large and messy data source. Web-scraping techniques allowed obtaining, cleaning, and filtering the data from the tourist social networks sites. Due to the volume, variety, and velocity features, Big Data pipelines were designed and implemented for data processing;
- Significance and value of the domain. NLP techniques were applied to filter the terms to build the knowledge base of the ontology;
- Ontology construction: Big Data provided facilities for the data preprocessing so that later, an ontological building tool facilitated the creation of the thesaurus, the

classifications, the taxonomy, the concept sets, the link between concepts, documentation, grouping in collections, mapping employing concept schemes, inference, and mapping link;

- The reasoning. The bidirectional relationship of Big Data semantics was fundamental in the application of the OntoTouTra ontology. The semantic basis was the ontology. For instance, we set axioms that determined the polarity of the tourist reviews.

The tourist reviews from the OTA gathered through web scraping became the ideal input to apply Big Data analytics because these fulfilled its features. The tourist reviews offered various aspects concerning the domain of the TTS, such as the location, time, services, ratings, and of course, the opinion. We extracted these aspects with opinion-mining techniques, and we had challenges such as the classification of multi-aspect opinions. We identified the vocabulary from these reviews using the NLP as a first step to face this challenge. We used machine learning classification methods and, in some cases, deep learning.

Besides the construction of the ontology, we also used Big Data analytics for its use, for instance, in the data visualizations such as Figure E.9 and the predictions such as the review scores depicted in the algorithm of Figure E.14. We used this algorithm for a double function: to generate the ontology vocabulary corpus and, in turn, to predict ratings. The algorithm preprocessed the data from the reviews. Specifically, in the data cleaning, we used the lemmatizing of the reviews and other NLP techniques such as tokenizing by word and by sentence, filtering stopwords, stemming, and tagging. To this end, we worked on Python libraries such as SpaCy, NLTK, Tokenizer, and Keras pad_sequences. We were also able to identify the language of the review. In the case of English, using this algorithm, we formed a vocabulary size of 16,466 terms and a maximum sentence length of 197 characters from a dataset of 57,063 instances. Each instance had a positive or negative review or both. The analysis of this vocabulary defined the classes with their attributes and their relationships. This definition was checked with the sources of the tourism authorities to enrich the definition of the ontology. The vocabulary was obtained in the sixth step, “Keras_create_vocabulary.” The remaining steps of the algorithm were intended to generate the model, train it, and predict the rating of the reviews as an application of the use of the ontology. Figure E.15 shows the validation results of the prediction of this algorithm in both loss and accuracy. Reasonable results can be seen, although low prediction. A model based on a bidirectional long short-term memory (LSTM) network and four fully connected layers was used. We could reduce the overfitting further by increasing the dropout layers of this deep-learning model.

5.4.13.4 Classification Using Big Data

Big Data analytics describes data, control technologies, analysis methods, and data mining development [162]. OntoTouTra’s data sources are ubiquitous, primarily social networks. We used data mining and Big Data analytics as a decision support process by searching raw data for hidden patterns that are useful and interpretable for decision-making in the TTS domain. In this way, we extracted facts and generated hypotheses using statistical tools, artificial intelligence, and machine learning.

We found the use of Big Data beneficial for processing structured, semistructured, and unstructured data (see Section 5.4.13.2) due to the web scraping applied to an OTA because data, especially tourist reviews, are characterized by the Big Data 3V requirement (volume, velocity, and variety).

The Big Data analytics applications for this study are synthesized as follows:

- Refinement of the ontology: A vocabulary was generated with NLP techniques (see Section 5.4.13.3) to obtain the glossary of the TTS domain to implement the stages of the specification and conceptualization of the ontology (see Section 5.4.4, Table 5.2);
- Data validation and cleaning: Using data-mining and text-mining techniques, we applied text preprocessing to the tourist reviews (see Sections 5.4.12 and 5.4.13.3 and Figure E.14), such as tokenization to obtain terms by removing spaces in blank and other punctuation symbols; removal of numbers so as not to affect the review sentiment measurement; elimination of stopwords; removal of scores; stemming according to language; and applying filters to determine the effect of a denial;
- Classification of reviews: The reviews provided us with different categories of data, and based on these categories, we were able to classify them. Not all categories were present in a review. Depending on the category, we applied supervised- and unsupervised-machine-learning classification algorithms. Table 5.7 depicts the categories identified in the reviews and the type of classification algorithm used depending on whether the reviews had labels;
- Prediction of reviews rating: We used a bidirectional-LSTM-network-based classifier to predict ratings using the vocabulary generated from the review terms (see Section 5.4.13.3 and Figure E.14);
- Data visualization: Using the programming and processing model, MapReduce, we generated Big Data datasets with a distributed and parallel algorithm on a cluster. We

used the map procedure to filter and sort the displayed data, and we executed the summary operations with the reduce method. An example is the heat map visualization in Figure E.9, where we mapped the country's regions and reduced the hotels count by region to represent them on a map with the `plotly.express` library.

Table 5.7 Tourist review categories.

Category	Classifier	Algorithm or Tool
Determine the polarity	Supervised	<code>nltk.sentiment.sentiment_analyzer</code>
Grouping by ratings	Not supervised	K-means
Detection of services	Supervised	Named entity recognition (NER) with SpaCy
Detection of tourist experiences	Supervised	NER with SpaCy
Detection of POIs	Supervised	NER with SpaCy
Detection of language	Supervised	<code>nltk.stem</code>

5.5 Making-decision system

With the ontology in place, this subsystem performs different level queries to retrieve the information that the DMO requires. The subsystem can carry out the queries from two alternatives: the traditional SPARQL queries and the HTTP request through API. By default, the queries are programmed through API since they are faster in execution because they are carried out in the backend of the system, and their results are reported in standard formats (JSON) that are compatible with the visualization and dashboard systems.

This subsystem has an alert component triggered by two main events: a non-conformity of the service or a performance report (see Figure 3.6). The non-conformity of the service is obtained through the reviews of the tourists processed in the opinion mining subsystem. Performance is processed based on previously defined KPI.

By triggering events, this subsystem generates traceability processes backward to find the origin of the issue and forward to determine the impact on the value chain of the tourism service (see Sections 3 and Appendix A).

The portfolio of tourist experiences, more than a report, is an open interface for communication with other systems that involve TTS through API that allow consultation on `OntoTouTra` [92].

Chapter 6

Results

6.1 Evaluation

6.1.1 Evaluation of the Ontology

In evaluating the ontology, we verified whether the objectives defined in the “purpose of the ontology” stage were met and verified whether the ontology was built correctly. We considered the quality criteria proposed by Gruber [163]: clarity, coherence, extensibility, minimal coding bias, and minimal ontological commitment, as the evaluative metrics of the ontology. First, we checked the internal consistency of the ontology; we used the HermiT reasoning [164] tool, included in Protégé. Once this reasoner was executed, no semantic, infinite loops or partition errors were found. As a second tool, we used OOPS! [165] to detect pitfalls in the ontology, which listed a minor pitfall related to the URI containing the file extension “.owl.” As a minor suggestion, we skipped this pitfall.

Then, we used the GQM approach of the FOCA methodology [166], consisting of the thirteen questions observed in Table 6.1. The objective of this approach is to verify the domain and application of the ontology.

Table 6.1 Applying the goal–question–metric approach from the FOCA methodology on the TTS ontology domain.

Goal	Question	Metric	Note	Question Grade	Goal Grade
1. Check if the ontology complies with substitutes	Q1. Were the competency questions defined?	Completeness	13 KPIs as CQ	100	83.3
	Q2. Were the competency questions answered?	Completeness	13 KPIs answered	100	
	Q3. Did the ontology reuse other ontologies?	Adaptability	Open link data with GeoNames and Time Ontology	50	
2. Check if the ontology complies with ontological commitments	Q4. Did the ontology impose a minimal ontological commitment?	Conciseness	Ontology uses abstractions to define concepts	75	75
	Q5. Did the ontology impose a maximum ontological commitment?	Conciseness	Ontology does not use many primitive concepts	-	
	Q6. Are the ontology properties coherent with the domain?	Consistency	Checked by Hermit reasoning (Protégé plugin)	75	
3. Check if the ontology complies with intelligent reasoning	Q7. Are there contradictory axioms?	Consistency	Checked by Hermit reasoning (Protégé plugin)	100	100
	Q8. Are there redundant axioms?	Conciseness	Checked by Hermit reasoning (Protégé plugin)	100	
4. Check if the ontology complies with efficient computation	Q9. Did the reasoner bring modeling errors?	Computational efficiency	1 minor error; Checked by OOPS!	75	75
	Q10. Did the reasoner perform quickly?	Computational efficiency	Depending on Protégé capacity (we ran without the reviews' individuals: 17.197 ms)	75	
5. Check if the ontology complies with human expression	Q11. Is the documentation consistent with modeling?	Clarity	Documentation generated by Protégé	100	100
	Q12. Were the concepts well written?	Clarity	We used the ontology annotations (rdfs:comment)	100	
	Q13. Are there annotations in the ontology that show the definitions of the concepts?	Clarity	We used the ontology annotations (rdfs:comment)	100	

The FOCA methodology is ideal for evaluating ontologies based on the GQM approach for an empirical evaluation, knowledge representation roles, and metrics based on the evaluation criteria. After iteration or in total, the GQM approach is executed, and finally, the quality of the ontology is calculated. First, the ontology validation must consider the type of ontology, whether a domain, task, or application. In the case of OntoTouTra, we think it is an application ontology because the concepts are described depending on a particular domain and task, in our case the TTS, which are specializations of related ontologies, as is the case of ontologies of the tourist domain.

FOCA considers criteria such as the clarity of the ontology, that is the definitions of concepts that arise from social situations. Another criterion is consistency, which guarantees that the ontology is consistent with its purposes. Completeness takes into account the whole meaning of individuals. On the other hand, adaptability refers to the reaction of the ontology to small changes in the axioms, and computational efficiency examines the ease and success by which reasoners can process the ontology.

Concerning the GQM approach, the objectives are defined in questions to extract information from the models. Moreover, the questions define a set of metrics for interpretation. In this way, the FOCA methodology raises five verification objectives. For each objective, a set of questions is posed (thirteen in total) that seek to interpret six metrics.

Regarding the last step of the FOCA methodology, the quality check, the evaluator verifies the questions and calculates their grades using the beta regression models proposed by Ferrari [167]. The authors of FOCA considered this model very appropriate since it is commonly used to model random varieties that assume values in the interval of the unit $(0, 1)$, such as rates, percentages, and proportions. The beta density can show different forms depending on the values of the parameters. Finally, it should be clarified that the authors recognized that there are questions with some degree of subjectivity, especially Questions 7 to 9, which can affect the final score; however, the regression model considers different weights for each of the parameters.

The ontology's quality was calculated by the beta regression models [167], as shown in Equation (6.1):

$$\begin{aligned}
 x &= -0.44 + 0.03(Cov_S \cdot Sb)_i + 0.02(Cov_C \cdot Co)_i + 0.01(Cov_R \cdot Re)_i \\
 &\quad + 0.02(Cov_{Cp} \cdot Cp)_i - 0.66 \cdot LEp_i - 25(0.1 \cdot Nl)_i \\
 \hat{\mu}_i &= \frac{\exp(x)}{1 + \exp(x)}
 \end{aligned} \tag{6.1}$$

where:

Cov_S = Goal 1 grade; Cov_C = Goal 2 grade; Cov_R = Goal 3 grade; Cov_{Cp} = Goal 4 grade; $LExp$ = experience of the evaluator; vast experience: $LExp$ is one, if not, zero; Nl = one only if some goal was impossible to answer all the questions; $Sb = 1, Co = 1, Re = 1, Cp = 1$ = because the total quality considers all the roles.

The equation, using the goal grades and considering that the evaluators have some experience, is:

$$\begin{aligned}
 x &= -0.44 + 0.03(83.3 \cdot 1) + 0.02(75.0 \cdot 1) + 0.01(100.0 \cdot 1) + 0.02(75.0 \cdot 1) - 0.66 \cdot 0 - 25(0.1 \cdot 0) \\
 \hat{\mu} &= \frac{\exp(6.059)}{1 + \exp(6.059)} \\
 \hat{\mu} &= 0.9977
 \end{aligned}
 \tag{6.2}$$

Thus, the total quality of the ontology is 99% (Equation (6.2)), which shows that the ontology's quality is high. Thus, OntoTouTra was successfully validated and verified.

6.1.2 Conceptual Validation

To validate the conceptual model, we used a set of tests applied to a use case to demonstrate the effectiveness of the OntoTouTra ontology using SPARQL queries. These tests were designed with an approach oriented toward the data of real cases gathered from one of the OTAs, using web-scraping techniques. The algorithm was executed with data from Colombia as a tourist destination, which was the selected use case. To answer the questions of the experts [168] in the TTS knowledge domain, we set some KPIs based on [169]. The indicators were grouped into four boxes: Satisfaction, Economy, Sustainability, and Organizational.

The KPIs are interpreted in the knowledge base as questions (CQ) that are answered through queries to the ontology. For each KPI, we developed test cases using SPARQL queries. We chose some KPIs from the document and adapted other indicators according to the TTS. We chose the ten most representative KPIs for the test, taking into account space-time variables in the queries. Furthermore, these queries can be broken down into different levels of grouping and detail, such as geographic areas, timelines, services, tourist experiences, and types of accommodation, among others. In the ten selected queries, we tried to involve these types of groupings in general detail. The selected KPIs are depicted in Table 6.2.

Table 6.2 KPI list.

Box	KPI	Indicator
1	01	% of visitors who rate the overall visitor experience as good or excellent
1	02	% of customers who consider the overall impression of the WiFi service to be good or excellent
2	03	Number of day visitors
3	04	Number of tourism enterprises (accommodation) per 10,000 population
3	05	Ratio of number of reviews to local population
3	06	Population rate with hotel influence
2	07	Foreign tourist arrivals (FTAs)
2	08	Inbound and domestic tourism
2	09	Seasonality patterns
2	10	Tourist experiences

6.1.3 Ontology Testing

The approach to using KPIs as test cases allowed evaluating the ontology from several indicators: semantics, inferences from ontological terms, consistency of the purpose of the ontology, and detection of inconsistencies. Table 6.2 depicts the test cases for each of the selected KPIs. As a reference for comparison, local government and WTO sources were sought to contrast the expected results (see Table 6.3). The test cases were run using SPARQL queries whose results demonstrated the reliability of the ontology when compared with the expected results (see the Appendix A). The execution of the test cases was performed with the Apache Jena Fuseki tool. The results are evidenced in this Supplementary Material.

In Table 6.3, we observe the results of the ontology test from a conceptual point of view, according to the application domain. The column “Test case” corresponds to the KPIs to validate. The column “Expected results” corresponds to the projected results after the test case (SPARQL query) has been executed. We compared these results with the sources, which are tourism authorities indicated in the column “Comparison sources.” From these sources, we identified the comparison data shown in the column “Source’s data.” We obtained results when executing the SPARQL queries, and these are listed in the column “Results obtained.” Based on these last two columns, we compared the consistency of the results. This comparison must be considered proportionally. The data from these sources were consolidated from the tourism sector, while the ontology data came from a portion of this sector that we obtained from the OTAs.

Table 6.3 Expected results.

Test Case	KPI	Expected Results	Comparison Sources	Source's Data	Results Obtained	Note
T001	1	Over 60 % of visitors rated the experience as good or excellent		-	71.56%	
T002	2	In Colombia, over 50% of customers considered the WiFi service to be good or excellent		-	53.5%	
T003	3	In Colombia, in 2019, over 1000 reviews per day	Colombia's Fact Sheets [170] pages 1–2	4,100,000 annual (2019)	2423 (mean)	Booking's reviewers represent the 21.57% visitors
T004	4	In Colombia, two (2) accommodation enterprises per 10,000 population	Colombia's Fact Sheets [170] page 4	5.6	2.33	28,000 establishments/50 million inhabitants = 5.6. Booking = 2.33
T005	5	The number of reviews depends on the local tourism industry (33 departments in Colombia)	[151] page 18	Bogotá, Antioquia, Bolívar	Bogotá, Antioquia, Bolívar	Top-3 departments
T006	6	Population rate with hotel influence depends on the local tourism industry	Colombia's Tourism Report [151] page 28	San Andrés, Bolívar, Bogotá	Bogotá, San Andrés, Valle	Top 3 departments
T007	7	Top 10 foreign tourist arrivals (FTAs) in Colombia	Colombia's Tourism Report [151] page 7	USA, Peru, France	USA, France, Argentina	Top 3 countries
T008	8	Inbound and domestic tourism in Colombia per department	Colombia's Fact Sheets [170] pages 1–2	4,100,000	459,322	Inbound travels
T009	9	Seasonality patterns per month of 2019 in Colombia	UNWTO Seasonality [171]	January–March, July–August	January–April, July–August	Peak seasons
T010	10	Top 10 Tourist experiences in Colombia		-	Beach, tours, game room	Top 3 tourist experiences

The results in Table 6.3 demonstrate the OntoTouTra ontology’s effectiveness in retrieving conceptual information from the TTS domain. All the proposed indicators were achieved through the SPARQL queries. In addition, the open architecture of this ontology allows the use of different tools and technologies to access data from the endpoint, such as Apache Fuseki, Apache Jena, Protégé, Open Link Virtuoso, Fuseki SOH (REST API), and OBAs. For this reason, the column “Note” describes the special comparison considerations for each test case.

6.1.4 Analysis of the Results

The objective of this work was to provide a knowledge base for the tourist traceability system. This knowledge base was built with input data from ubiquitous data, mainly social networks, such as OTAs. This paper indicates the method to construct an ontology whose data sources are typical in Big Data environments. The features of the developed ontology called OntoTouTra are depicted in Table 6.4. In the Appendix A, we show the screenshots running OntoTouTra on each of these tools.

Table 6.4 OntoTouTra features.

Item	Feature	Tool
1	SPARQL Interface	Apache Jena Apache Jena Fuseki Protégé OpenLink Virtuoso
2	Web interface	RDFLib/Dash WebVOWL/TikZ [172]
3	REST API	Fuseki SOH Ontology-Based API (OBA)
4	Documentation	Protégé OBA

Table 6.5 summarizes the differences between OntoTouTra and the similar ontologies within the tourism domain, based on the studies of [173, 174]. Each ontology has its specific purpose within the field of tourism. For its development, common standards were used to generate the axioms. The number of concepts depends on the domain contemplated. When evaluating the ontology with use cases based on the KPIs, it was a challenge that we overcame when performing complex SPARQL queries, especially in the space–time dimensions that are sensitive in a TTS.

Table 6.5 OntoTouTra vs. other tourism ontologies.

Item	Domain	Use	Axioms
OntoTouTra	Tourist traceability	Decision-making at the destination	OWL
Mondeca	Tourism	Tourism concepts	OWL
HarmoNET	Tourism	Accommodation	OWL
Travel Itinerary	Travel	Tourist itineraries	OWL
Hontology	Hotel	Hotels	OWL
OnTour Project	e-Tourism	Accommodation	OWL
COTRIN	Open Travel Alliance (OTA) specifications	Travel industry	XML schema
LA_DMS project	DMO	Tourist destination	OWL-S
Hi-Touch project	Tourism products	Customer's expectations	OWL
TAGA	Travel concepts	Simulations	OWL

In this study, we used FOCA [166] as an ontology evaluation method because it allowed us to evaluate multiple quality criteria, which were the criteria based on Gruber's proposal [163] and served as the metrics of evaluation. Following FOCA and the beta regression modeling equation [167], the total quality was calculated based on the weights of each metric of the evaluation goals. In this way, a total quality score was obtained for the OntoTouTra ontology, taking into account the TTS domain, of 99.77%, indicating that the quality was high and satisfied the requirements of its domain. To achieve greater objectivity in this weighting, we used ontology evaluation tools such as HermiT [164] and OOPS! [165]. The first one allowed us to provide the reasoning for the consistency of the content of the ontology, and the second one detected the pitfalls. The results generated by both tools were satisfactory.

Chapter 7

Conclusions and future work

In this research, different alternatives for tourist traceability were analyzed; through a systematic literature review, studies were grouped into two aspects: the tracking of tourists and the trajectories in the destination. Concerning tracking, the studies grouped under this topic describe tracking the flow of tourists in a single dimension: spatiality. In comparison, the studies grouped in the definition of trajectories in the destination contemplate time-space causality. However, a **TTS**, apart from these two components, involves the relationship and behavior of the tourist with the service in the different stages of the tourism value chain. Studies that fully comply with these three guidelines are not found in the literature review; a model was developed as an alternative to a **TTS** in this study.

In this sense, to develop this model, it was necessary to complete the studies carried out by Chantre et al. [103], who proposed the tourist traceability concept and determined its importance within the tourism ecosystem. For this reason, it was necessary to create a conceptual framework for tourist traceability based on international traceability standards and authorities in the tourism domain (see Section 3). Based on this conceptual framework, the **TTS** model was defined.

Regarding **OntoTouTra**, we proposed a model for building an ontology of a **TTS**, answering the research question “How can we develop a tourist traceability ontology based on gathering and processing ubiquitous data, using Big Data techniques?” The gap demonstrated in the state-of-the-art showed us the lack of an ontology whose domain was tourist traceability. Therefore, we proposed a model for the creation of the **OntoTouTra** ontology. In turn, we adapted the lifecycle of Big Data analytics presented by Erl et al. [4] to deal with the volume, variety, and velocity of data coming from ubiquitous sources, in particular from an **OTA**.

We applied the GQM approach of the FOCA methodology to validate the *OntoTouTra* ontology and achieved a score of 99.77% of the total quality of the ontology. We used *HermiT*, *Protégé*, and *OOPS!* as evaluation tools. However, the number of individuals in the ontology, especially tourist reviews, required enormous computational resources. For instance, we used *HermiT* as a *Protégé* [153] plugin, and the capacity of this tool restricted its execution. For the evaluation tests, we had to ignore the individuals of the tourist reviews. A new research challenge arises to adapt this type of ontological tool to Big Data environments. The amount of knowledge affects the quality of the ontology's testing processes, which is imperative in this environment.

The analysis of the ontology validation results demonstrated its functionality. The validation was conceptual, whose aim was to evaluate the purpose and functionality of the ontology. This goal was achieved by executing *SPARQL* queries for 10 *KPI* representative of a *TTS*.

As contributions of this study, we highlight the construction model of the ontology, the adaptation of the lifecycle of Big Data analytics so that the ontology works with ubiquitous data sources in Big Data contexts, and the interoperability of the ontology with open systems, since it allows *SPARQL* queries and *RESTFUL API*. The source code allowed the creation, access, and use of the ontology in Big Data environments, using *PySpark*, and the provision of the ontology for open link data, in particular with *GeoNames* and *Time Ontology*. The results of this study are a meaningful contribution to the scientific community and to *DMO* looking for a knowledge base to support decision-making regarding destination management.

The practical application of the developed ontology is extensive: it serves as a knowledge base to support decision-making in the destination, recommendation systems for tourist experiences, monitoring of the management of the *DMO*, the design or improvement of tourist experiences, the benchmarking of tourist experiences, tourist service providers, and web portals on destination tourist information, among others.

Through the *OntoTouTra* ontology, we plan to consolidate the knowledge base for *DMO*. As future work, we will include other ubiquitous computing sources, such as data from tourist mobile devices and sensors from *POI*. Besides, we will offer a portfolio of tourist experiences of the destination.

As contributions to this project, we can mention the following:

A knowledge base for the tourist traceability system based on an ontology called *OntoTouTra*. The ontology was built in Big Data environments because its data sources are ubiquitous: social networks, apps installed on tourists' mobile devices, and sensors located at *POI*. The ontology was built using the *METHONTOLOGY* methodology, and as a scientific contribution, the Big Data life cycle was adapted for the ontological construction.

On the other hand, the ontology directly interacting with the other components of the tourist traceability system, especially the decision-making component, can generate new information entities, such as classes or class attributes, according to the preprocessed data. The web scraping source code is another contribution of this research; it collects the data from an [OTA](#). As a case study, we worked with [Booking.com](#). This application receives the name of the geographical region as an input parameter. It automatically maps the underlying sub-regions to the minimum level that the [OTA](#) has of that region. Subsequently, for each destination obtained, it obtains the tourist properties of the destination, and for each property, data is collected on the property, its services, ratings, and tourist reviews. The web scraping application has an interface to provide this data structured to the [OntoTouTra](#) ontology.

Due to the characteristics of the data obtained through the web scraping application, especially the volume, variety, and velocity, it became necessary to support the application of Web scraping with a Big Data pipeline, which is one of the engineering contributions to the project. The Big Data pipeline processes the data in the form of streaming, and datasets and queries are consolidated on these streams, also in streaming. Grouping operations are performed using Map Reduce.

Although ontology is the backbone of the tourism traceability system, it is not the only component. In a symbiotic relationship to [OntoTouTra](#) are the opinion mining, location intelligence, and decision-making components. Opinion mining, using [NLP](#) techniques, processes the ubiquitous data coming from social networks. The location intelligence component calculates distances, proximity, and geographic trajectories of tourists and geographic entities in a tourism ecosystem. Finally, the decision-making component controls the ontological system (belief change) for its evolution and the population and enrichment of [OntoTouTra](#). In addition, through queries to the ontology that can be by [SPARQL](#) or through [API](#), it obtains information for decision making, which can be visually displayed geographically or through dashboards (by programming [KPI](#)). This last component can also generate portfolios of tourism experiences from the tourist's perspective. The Tourist Traceability System is the main contribution of the research, both for the domain of tourism and engineering and knowledge reuse issues.

At the time of ontological construction, a new contribution emerges from this research: the conceptual framework of tourism traceability. This framework aims to cover theoretical gaps on tourism traceability found when preparing this research's literature review. It is a contribution put to the consideration of the scientific community of the tourist domain. Thanks to this conceptual framework, the conceptual conception of [OntoTouTra](#) was defined and refined.

Publishing papers in recognized scientific journals carried out the socialization of these contributions. The results of the [OntoTouTra](#) work were published in Applied Sciences [142], the review of the state of the art of this research and other collaborative research (both investigations belonging to the same call MinCiencias 733) is published in Future Internet [92]. The processing of location data from sensors and tourist mobile device applications, together with the same collaborative research, is published in the Sensors paper [175]. Finally, the cooperative paper published in IEEE Access was an input for this research project, especially state-of-the-art deep learning algorithms [176].

The ontology and its documentation are published in the <http://tourdata.org/> repository, and the datasets and source code are in Github repositories.

Chapter 8

List of Acronyms

The following abbreviations are used in this document:

API	Application Programming Interfaces
CQ	Competency Questions
DMO	Destination Management Organizations
eWOM	electronic Word-of-Mouth
IoT	Internet of Things
ISO	International Organization for Standardization
KM	Knowledge Management
KPI	Key Performance Indicators
NLP	Natural Language Processing
OntoTouTra	Ontology for Tourist Traceability
OTA	Online Travel Agencies
OWL	Web Ontology Language
POI	Point of Interest
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SPARQL	SPARQL Protocol and RDF Query Language
TTS	Tourist Traceability System
UNWTO	United Nations - World Tourism Organization

References

- [1] K. Weiermair, “Prospects for Innovation in Tourism,” *Journal of Quality Assurance in Hospitality & Tourism*, vol. 6, pp. 59–72, 01 2005.
- [2] H. Song, J. Liu, and G. Chen, “Tourism Value Chain Governance: Review and Prospects,” *Journal of Travel Research*, vol. 52, pp. 15–28, 01 2013.
- [3] X. Zhang, H. Song, and G. Q. Huang, “Tourism Supply Chain Management: A New Research Agenda,” *Tourism Management*, vol. 30, pp. 345–358, 06 2009.
- [4] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concepts, Drivers & Techniques*, M. Taub, Ed. Prentice Hall, 2016.
- [5] S. Lohmann, S. Negru, F. Haag, and T. Ertl, “Visualizing ontologies with VOWL,” *Semantic Web*, vol. 7, no. 4, pp. 399–419, 2016. [Online]. Available: <http://dx.doi.org/10.3233/SW-150200>
- [6] H. Boley and E. Chang, “Digital ecosystems: Principles and semantics,” *IEEE International Conference on Digital Ecosystems and Technologies*, 2 2007.
- [7] A. Scharl, L. Lalicic, and I. Önder, “Tourism intelligence and visual media analytics for destination management organizations,” *Analytics in Smart Tourism Design Concepts and Methods*, pp. 165–178, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-44263-1_10
- [8] A. Scharl, A. Hubmann-Haidvogel, A. Weichselbraun, and H.-P. Lang, “From web intelligence to knowledge co-creation: A platform for analyzing and supporting stakeholder communication,” *IEEE Internet Computing*, vol. 15, no. 5, pp. 21–29, 9 2013. [Online]. Available: <http://ieeexplore.ieee.org/abstract/document/6527880/>
- [9] Z. Xiang and D. R. Fesenmaier, *Analytics in Smart Tourism: Design Concepts and Methods*, ser. Tourism on the Verge, P. J. Sheldon and

- D. R. Fesenmaier, Eds. USA: Springer, 2017. [Online]. Available: <http://www.springer.com/in/book/9783319442624>
- [10] J. L. Nicolau, "Travel demand modeling with behavioral data," *Analytics in Smart Tourism Design Concepts and Methods*, pp. 31–43, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-44263-1_3
- [11] M. Fuchs, A. Abadzhiev, B. Svensson, W. Höpken, and M. Lexhagen, "A knowledge destination framework for tourism sustainability: A business intelligence application from sweden," *Tourism : An International Interdisciplinary Journal*, vol. 61, no. 2, pp. 121–148, 7 2013. [Online]. Available: <http://hrcak.srce.hr/106864?lang=en>
- [12] H. Song and H. Liu, "Predicting tourist demand using big data," *Analytics in Smart Tourism Design Concepts and Methods*, pp. 13–30, 2017.
- [13] J. J. Kim and D. R. Fesenmaier, "Measuring human senses and the touristic experience: Methods and applications," *Analytics in Smart Tourism Design Concepts and Methods*, pp. 47–64, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-44263-1_4
- [14] E. Cohen, "A phenomenology of tourist experience," *Sociology*, vol. 13, no. 2, pp. 179–201, 5 1979. [Online]. Available: https://www.researchgate.net/publication/249824872_A_Phenomenology_of_Tourist_Experience
- [15] Y. Choe and D. R. Fesenmaier, "The quantified traveler: Implications for smart tourism development," *Analytics in Smart Tourism Design Concepts and Methods*, pp. 65–77, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-44263-1_5
- [16] D. Li and Y. Yang, "Gis monitoring of traveler flows based on big data," *Analytics in Smart Tourism Design Concepts and Methods*, pp. 111–126, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-44263-1_7
- [17] M. Scaglione and J.-P. Trabichet, "Using mobile data and strategic tourism flows. pilot study monitour in switzerland," *Big Data & Business Intelligence in the Travel & Tourism Domain*, pp. 69–72, 4 2016. [Online]. Available: https://www.researchgate.net/publication/305963568_Using_Mobile_data_and_strategic_tourism_flows_Pilot_study_Monitour_in_Switzerland

- [18] C. Demunter, *Tourism statistics in the European Statistical System 2008 data*, Eurostat, Ed. Eurostat, 2010. [Online]. Available: <http://ec.europa.eu/eurostat/documents/3888793/5848705/KS-RA-10-010-EN.PDF/cb99b249-57f6-4c8d-8a5f-573982c9fbd1>
- [19] The Council of the European Union, *Council Directive 95/57/EC on the collection of statistical information in the field of tourism*, Official Journal of the European Communities, The Council of the European Union Std. L 291 /32, 12 1995. [Online]. Available: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0057&qid=1497904282670&from=en>
- [20] European Parliament and of the Council, *Regulation (EU) No 692/2011 of the European Parliament and of the Council*, Official Journal of the European Communities, European Parliament and of the Council Std. L 192/17, 7 2011. [Online]. Available: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R0692&from=EN>
- [21] U. Hansmann, L. Merk, M. S. Nicklous, and T. Stober, *Pervasive Computing Handbook*, 1st ed., I. D. E. GmbH, Ed. Springer, 2001.
- [22] J. Krumm, *Ubiquitous Computing Fundamentals*, 1st ed., J. Krumm, Ed. Chapman & Hall/CRC, 2010.
- [23] M. Weiser, “The computer for the 21st century,” *Scientific American*, 94-104 1991. [Online]. Available: <https://www.lri.fr/~mbl/Stanford/CS477/papers/Weiser-SciAm.pdf>
- [24] F. Mattern, “Visión y fundamentos técnicos de la computación ubicua,” *Novatica / Upgrade*, no. 153, pp. 4–7, 9 2001.
- [25] M. Mühlhäuser, “Introduction to ubiquitous computing,” *Handbook of Research: Ubiquitous Computing Technology for Real Time Enterprises*, 2007.
- [26] M. Satyanarayanan, “Pervasive computing: Vision and challenges,” *IEEE Personal Communications*, pp. 1–10, 2001. [Online]. Available: <https://www.cs.cmu.edu/~aura/docdir/pcs01.pdf>
- [27] S. Poslad, *Ubiquitous Computing Smart Devices, Environments and Interactions*. A John Wiley and Sons, Ltd, Publication, 2009.

- [28] E. Marchiori and L. Cantoni, "Evaluating destination communications on the internet," *Analytics in Smart Tourism Design Concepts and Methods*, pp. 253–280, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-44263-1_15
- [29] D. Buhalis and S. H. Jun, "E-tourism," *CTR Contemporary Tourism Reviews*, pp. 1–38, 2011.
- [30] B. Neuhofer, D. Buhalis, and A. Ladkin, "Smart technologies for personalized experiences: a case study in the hospitality domain," *Electron Markets*, pp. 1–12, 2 2015. [Online]. Available: https://www.academia.edu/10672841/Neuhofer_B._Buhalis_D._and_Ladkin_A._2015_Smart_technologies_for_personalized_experiences_a_case_study_in_the_hospitality_domain
- [31] J. Cleland-Huang, O. Gotel, and A. Zisman, *Software and Systems Traceability*, 1st ed., O. G. Jane Cleland-Huang and A. Zisman, Eds. Springer, 2012.
- [32] GS1, *GS1 Standards Document GS1 Global Traceability Standard*, GS1 (previously: The European Article Numbering (EAN) Association) Std., 11 2012. [Online]. Available: http://www.gs1.org/sites/default/files/docs/traceability/Global_Traceability_Standard.pdf
- [33] J. H. Kang and G. Borriello, "Ubiquitous computing using wireless broadcast," *IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 2004)*, vol. 6ed, 2004. [Online]. Available: <http://ieeexplore.ieee.org/document/1377316/?part=1>
- [34] J. M. Rickly and S. McCabe, "Authenticity for tourism design and experience," *Design Science in Tourism Foundations of Destination Management*, pp. 55–68, 2017. [Online]. Available: https://www.researchgate.net/publication/311999586_Authenticity_for_Tourism_Design_and_Experience
- [35] D. R. Fesenmaier and Z. Xiang, *Design Science in Tourism Foundations of Destination Management*, ser. Tourism on the Verge, D. R. Fesenmaier and Z. Xiang, Eds. Springer, 2017.
- [36] I. P. Tussyadiah, "Technology and behavioral design in tourism," *Design Science in Tourism Foundations of Destination Management*, 2017. [Online]. Available: https://link.springer.com/chapter/10.1007%2F978-3-319-42773-7_12
- [37] F. J. Zach and D. Krizaj, "Experiences through design and innovation along touch points," *Design Science in Tourism: Foundations of Destination Management*, pp. 215–232, 2017.

- [38] I. P. Tussyadiah, "Toward a theoretical foundation for experience design in tourism," *Journal of Travel Research*, vol. 53, pp. 543–564, 7 2014. [Online]. Available: <http://journals.sagepub.com/doi/abs/10.1177/0047287513513172>
- [39] G. Moscardo, "Stories as a tourist experience design tool," *Design Science in Tourism Foundations of Destination Management*, pp. 97–124, 2017.
- [40] R. Ek, J. Larsen, S. Buhl Hornskov, and O. Kjaer Mansfeldt, "A Dynamic Framework of Tourist Experiences: Space-Time and Performances in the Experience Economy," *Scandinavian Journal of Hospitality and Tourism*, vol. 8, no. 2, pp. 122–140, 2008.
- [41] L. Mossberg, "A marketing approach to the tourist experience," *Scandinavian Journal of Hospitality and Tourism*, vol. 7, no. 1, pp. 59–74, 2007.
- [42] Corporación para el Desarrollo Turístico - Codetur, "Aplicaciones móviles de los destinos turísticos españoles. informe de investigación 2013," Proyecto Codetur, Tech. Rep., 10 2013. [Online]. Available: http://invattur.aimplas.es/ficheros/noticias/107090748E70_01_informe_apps_codetur_oct2013.pdf
- [43] M. R. Ebling, "Pervasive computing and the internet of things," *IEEE Pervasive Computing*, vol. 15, no. 1, pp. 2–4, 1 2016. [Online]. Available: <http://ieeexplore.ieee.org/document/7389283/>
- [44] J. L. V. Barbosa, "Ubiquitous computing: Applications and research opportunities," *IEEEExplore*, 12 2015. [Online]. Available: <http://ieeexplore.ieee.org/stamp/stamp.jsp?reload=true&tp=&arnumber=7435625>
- [45] United Nations, *Data protection regulations and international data flows: Implications for trade and development*, U. Nations, Ed. UNCTAD, 2016. [Online]. Available: http://unctad.org/en/PublicationsLibrary/dtlstict2016d1_en.pdf
- [46] N. Cuculeski, I. Petrovska, and T. Petkovska Mircevska, "Emerging trends in tourism: Need for alternative forms in macedonian tourism," *Review of Innovation and Competitiveness (RIC)*, vol. 1, no. 1, pp. 103–114, 11 2015. [Online]. Available: <http://hrcak.srce.hr/155599?lang=en>
- [47] U. Gretzel, M. Sigala, Z. Xiang, and C. Koo, "Smart tourism: foundations and developments," *Electron Markets*, 8 2015. [Online]. Available: https://www.researchgate.net/publication/280719315_Smart_tourism_foundations_and_developments

- [48] A. Mooghali, R. Alijani, N. Karami, and A. Khasseh, "Scientometric analysis of the scientometric literature," *International Journal of Information Science and Management*, vol. 9, 01 2011.
- [49] J. Ruiz-Rosero, G. Ramirez-Gonzalez, and R. Khanna, "Field programmable gate array applications—a scientometric review," *Computation*, vol. 7, no. 4, 2019. [Online]. Available: <https://www.mdpi.com/2079-3197/7/4/63>
- [50] United Nations Global Compact Office, *A Guide to Traceability*. BSR, 2014. [Online]. Available: https://www.unglobalcompact.org/docs/issues_doc/supply_chain/Traceability/Guide_to_Traceability.pdf
- [51] International Organization for Standardization, *ISO 12875:2011. Traceability of finfish products*. BSI, 2011. [Online]. Available: <https://www.iso.org>
- [52] B. McKercher and G. Lau, "Movement patterns of tourists within a destination," *TOURISM GEOGRAPHIES*, vol. 10, pp. 355–374, 2008.
- [53] Çetinkaya, C. and Kabak, M. and Erbas, M. and Özceylan, E., "Evaluation of ecotourism sites: a GIS-based multi-criteria decision analysis," *Kybernetes*, vol. 47, 04 2018.
- [54] B. McKercher, N. Shoval, E. Ng, and A. Birenboim, "First and repeat visitor behaviour: Gps tracking and gis analysis in hong kong," *TOURISM GEOGRAPHIES*, vol. 14, pp. 147–161, 2012.
- [55] B. McKercher, N. Shoval, E. Park, and A. Kahani, "The [limited] impact of weather on tourist behavior in an urban destination," *JOURNAL OF TRAVEL RESEARCH*, vol. 54, pp. 442–455, 2015.
- [56] Y. Yang and K. K. F. Wong, "Spatial distribution of tourist flows to china's cities," *TOURISM GEOGRAPHIES*, vol. 15, pp. 338–363, 2013.
- [57] F. L. Miller, "Using a gis in market analysis for a tourism-dependent retailer in the pocono mountains," *JOURNAL OF TRAVEL & TOURISM MARKETING*, vol. 25, pp. 325–340, 2008.
- [58] J. Ren, "Assessing the carrying capacity of tourist resorts: An application of tourists' spatial behavior simulator based on GIS and multi-agent system," *Wuhan University Journal of Natural Sciences*, 2004.

- [59] E. Ardizzone, M. F. Di, C. M. La, and G. Mazzola, "Extracting touristic information from online image collections," in *Signal Image Technology and Internet Based Systems*, 2012, pp. 482–488, in Proceedings of the 8th International Conference on Signal Image Technology and Internet Based Systems, SITIS 2012, Sorrento, 25 November 2012 through 29 November 2012.
- [60] M. Deidda, A. Pala, and G. Vacca, "A tourist Location Based Service (LBS) for the Cagliari city," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, vol. 38, 01 2010.
- [61] Q. Huang, L. Xia, and D. Wu, "An enhanced hybrid lbs and its prototype for supporting backpackers in beijing," *WIRELESS PERSONAL COMMUNICATIONS*, vol. 77, pp. 433–448, 2014.
- [62] M. Ferrante, S. De Cantis, and N. Shoval, "A general framework for collecting and analysing the tracking data of cruise passengers at the destination," *CURRENT ISSUES IN TOURISM*, vol. 21, pp. 1426–1451, 2018.
- [63] A. Domenech, A. Gutierrez, and S. Anton Clave, "Cruise passengers' spatial behaviour and expenditure levels at destination," *TOURISM PLANNING & DEVELOPMENT*, vol. 17, pp. 17–36, 2020.
- [64] Y. Li, L. Yang, H. Shen, and Z. Wu, "Modeling intra-destination travel behavior of tourists through spatio-temporal analysis," *Journal of Destination Marketing and Management*, vol. 11, pp. 260–269, 2019.
- [65] A. Y. Grinberger and N. Shoval, "Spatiotemporal contingencies in tourists' intradiurnal mobility patterns," *JOURNAL OF TRAVEL RESEARCH*, vol. 58, pp. 512–530, 2019.
- [66] N. Gali and S. Aulet, "Tourists' space-time behavior in heritage places: Comparing guided and nonguided visitors," *INTERNATIONAL JOURNAL OF TOURISM RESEARCH*, vol. 21, pp. 388–399, 2019.
- [67] B. Mckercher, A. Hardy, and J. Aryal, "Using tracking technology to improve marketing: insights from a historic town in tasmania, australia," *JOURNAL OF TRAVEL & TOURISM MARKETING*, vol. 36, pp. 823–834, 2019.
- [68] M. Sowmya, S. Prakash, S. Singh, S. Maloo, and S. Yadav, "Smart tourist guide (touristo)," in *Emerging Research in Computing, Information, Communication and Applications*, vol. 906. Springer Verlag, 2019, pp. 299–312, in Proceedings of the

- 5th International Conference on Emerging Research in Computing, Information, Communication and Applications, ERCICA 2018, 27 July 2018 through 28 July 2018.
- [69] E. Regkou and M. Dasygenis, "Design and development of an android application with supportive website in order to create a travel guide for Western Macedonia," in *Proceedings of the 22nd Pan-Hellenic Conference on Informatics*. Association for Computing Machinery, 2018, pp. 168–173, in Proceedings of the 22nd Pan-Hellenic Conference on Informatics, PCI 2018, 29 November 2018 through 1 December 2018.
- [70] J. Raun, N. Shoval, and M. Tiru, "Gateways for intra-national tourism flows: measured using two types of tracking technologies," *International Journal of Tourism Cities*, vol. 6, pp. 261–278, 2020.
- [71] G. K. Riungu, B. A. Peterson, J. A. Beeco, and G. Brown, "Understanding visitors' spatial behavior: a review of spatial applications in parks," *TOURISM GEOGRAPHIES*, vol. 20, pp. 833–857, 2018.
- [72] S.-Y. Yang and C.-L. Hsu, "A location-based services and google maps-based information master system for tour guiding," *COMPUTERS & ELECTRICAL ENGINEERING*, vol. 54, pp. 87–105, 2016.
- [73] J. Hendrikx, J. Johnson, and C. Shelly, "Using gps tracking to explore terrain preferences of heli-ski guides," *JOURNAL OF OUTDOOR RECREATION AND TOURISM-RESEARCH PLANNING AND MANAGEMENT*, vol. 13, pp. 34–43, 2016.
- [74] N. Shoval and R. Ahas, "The use of tracking technologies in tourism research: the first decade," *Tourism Geographies*, vol. 18, pp. 1–20, 08 2016.
- [75] Z. Xiang, U. Gretzel, and D. Fesenmaier, "Semantic representation of tourism on the internet," *Journal of Travel Research*, vol. 47, 05 2009.
- [76] J. Tribe and J. J. Liburd, "The tourism knowledge system," *Annals of Tourism Research*, vol. 57, pp. 44–61, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S016073831500170X>
- [77] S. Mouhim, A. aoufi, C. Cherkaoui, D. Hassan, and D. Mammass, "A knowledge management approach based on ontologies: The case of tourism," *International journal of computer science and emerging technologies*, vol. 2, 12 2011.
- [78] M. Uschold and M. Grüninger, "Ontologies: Principles, methods and applications," *The Knowledge Engineering Review*, vol. 11, 01 1996.

- [79] M. Missikoff and F. Taglino, *An Ontology-based Platform for Semantic Interoperability*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 617–633.
- [80] O. Carloni, “Boolean formulas of simple conceptual graphs SGBF,” in *In Proceedings of the Second International Conference on Graph Structures for Knowledge Representation and Reasoning*, 07 2011, pp. 18–67.
- [81] K. Siorpaes and D. Bachlechner, “Ontour: Tourism information retrieval based on yars,” *ESWC 2006*, 2006.
- [82] K. Prantner, Y. Ding, M. Luger, Z. Yan, and C. Herzog, “Tourism ontology and semantic management system: State-of-the-arts analysis,” *IADIS International Conference: IADIS*, 01 2007.
- [83] W. V. Siricharoen, “Using Ontologies for E-tourism,” in *In Proceedings of the 4th WSEAS/IASME International Conference on Engineering Education (EE 2007) Proceeding*, 2007.
- [84] X. Zhao, L. Liu, H. Wang, and W. Song, “Ontology Construction of the Field of Tourism in Africa,” in *In Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, 12 2015, pp. 47–50.
- [85] Y. Huang and L. Bian, “Using ontologies and formal concept analysis to integrate heterogeneous tourism information,” *IEEE Transactions on Emerging Topics in Computing*, vol. 3, pp. 1–1, 06 2015.
- [86] A. Valls, K. Gibert, A. Orellana, and S. Antón-Clavé, “Using ontology-based clustering to understand the push and pull factors for british tourists visiting a mediterranean coastal destination,” *Information & Management*, vol. 55, no. 2, pp. 145–159, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378720617303920>
- [87] M. R. Islam, B. A. Hossain, M. N. Imteaj, S. Akhter, H. S. Jogesh, and M. B. Mostafa, “Ontranetbd: A knowledgebase for the travel network in bangladesh,” in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017, pp. 170–174.
- [88] F. Giunchiglia and B. Dutta, “DERA: A Faceted Knowledge Organization Framework,” in *In Proceedings of the International Conference on Theory and Practice of Digital Libraries*, 2011.

- [89] F. Suchanek, G. Kasneci, and G. Weikum, “Yago - a large ontology from wikipedia and wordnet,” *Web Semantics : Science, Services and Agents on the World Wide Web, v.6, 203-217 (2008)*, vol. 6, 09 2008.
- [90] M. Rodríguez-García, R. Valencia-García, F. Garcia-Sanchez, and J. J. Samper Zapater, “Creating a semantically-enhanced cloud services environment through ontology evolution,” *Future Generation Computer Systems*, vol. 32, p. 295–306, 03 2014.
- [91] J. Llorens, J. Morato, G. Génova, J. Fuentes, V. Quintana, and I. Díaz, “Rhsp: an information representation model based on relationship,” *Studies in Fuzziness and Soft Computing*, vol. 159, pp. 221–253, 01 2004.
- [92] L. Santamaria-Granados, J. F. Mendoza-Moreno, and G. Ramirez-Gonzalez, “Tourist Recommender Systems Based on Emotion Recognition—A Scientometric Review,” *Future Internet*, vol. 13, no. 1, 2021. [Online]. Available: <https://www.mdpi.com/1999-5903/13/1/2>
- [93] Y. Chu, H. Wang, L. Zheng, Z. Wang, and K.-L. Tan, “TRSO: A Tourism Recommender System Based on Ontology,” in *In Proceedings of the International Conference on Knowledge Science, Engineering and Management*, vol. 9983, 10 2016.
- [94] H.-E. Guergour and Z. Boufaïda, “A domain ontology building process based on principles of social web,” in *2012 International Conference on Information Technology and e-Services*, March 2012, pp. 1–6.
- [95] A. Moreno, A. Valls, D. Isern, L. Marin, and J. Borràs, “Sigtur/e-destination: Ontology-based personalized recommendation of tourism and leisure activities,” *Engineering Applications of Artificial Intelligence*, vol. 26, pp. 633–651, 01 2013.
- [96] F. Girardin, F. Calabrese, F. Dal Fiore, C. Ratti, and J. Blat, “Digital footprinting: Uncovering tourists with user-generated content,” *Pervasive Computing, IEEE*, vol. 7, pp. 36 – 43, 01 2009.
- [97] M. Mariani and M. Borghi, “Effects of the booking.com rating system: Bringing hotel class into the picture,” *Tourism Management*, vol. 66, pp. 47–52, 06 2018.
- [98] V. Lytvyn, V. Vysotska, Y. Burov, and A. Demchuk, “Architectural Ontology Designed for Intellectual Analysis of E-Tourism Resources,” in *2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, vol. 1, 2018, pp. 335–338.

- [99] C. Lee, T.-C. Hsia, H.-C. Hsu, and J. Lin, “Ontology-based tourism recommendation system,” in *In Proceedings of the 2017 4th International Conference on Industrial Engineering and Applications (ICIEA)*, 04 2017, pp. 376–379.
- [100] A. Smirnov, A. Ponomarev, N. Shilov, A. Kashevnik, and N. Teslya, “Ontology-based human-computer cloud for decision support: Architecture and applications in tourism,” *International Journal of Embedded and Real-Time Communication Systems*, vol. 9, pp. 1–19, 01 2018.
- [101] G. Prasamuvarso Kuntarto, I. Gunawan, F. Moechtar, Y. Ahmadin, and B. I. Santoso, “Dwipa ontology iii: Implementation of ontology method enrichment on tourism domain,” *International Journal on Smart Sensing and Intelligent Systems*, vol. 10, pp. 903–919, 12 2017.
- [102] J. Borràs, J. Flor, Y. Perez, A. Moreno, A. Valls, D. Isern, A. Orellana, A. Russo, and S. Clavé, “SigTur/E-Destination: A System for the Management of Complex Tourist Regions,” in *Information and Communication Technologies in Tourism*, 01 2011, pp. 39–50.
- [103] A. Chantre Astaiza, L. Fuentes-Moraleda, A. Muñoz-Mazón, and G. Ramirez-Gonzalez, “Science mapping of tourist mobility 1980–2019. technological advancements in the collection of the data for tourist traceability,” *Sustainability*, vol. 11, p. 4738, 08 2019.
- [104] M. Wick, “Geonames Ontology,” Unxos GmbH: Wollerau, Switzerland, Tech. Rep., 2015. [Online]. Available: <http://www.geonames.org/about.html>
- [105] F. Frontini, R. Del Gratta, and M. Monachini, “GeoDomainWordNet: Linking the Geonames Ontology to WordNet,” in *In Proceedings of the Language and Technology Conference*, vol. 9561, 07 2016, pp. 229 – 242.
- [106] G. Team. (2019) GeoNames Webservice Subdivision Levels. [Online]. Available: <https://www.geonames.org/export/subdiv-level.html>
- [107] DANE. (2019) Geovisor de Consulta de Codificación de la División. [Online]. Available: <https://geoportal.dane.gov.co/geovisores/territorio/consulta-divipola-division-politico-administrativa-de-colombia/>
- [108] S. Cox and C. Little. (2016) Time Ontology in Owl. [Online]. Available: <https://www.w3.org/TR/owl-time/>

- [109] International Open Data Charter ODC. (2019) ODC Principles. [Online]. Available: <https://opendatacharter.net/principles/>
- [110] Ministerio de Tecnologías de la Información y las Comunicaciones. (2019) Datos Abiertos. [Online]. Available: <https://www.datos.gov.co/>
- [111] Situr Boyacá. (2019) Sistema de Información Turística de Boyacá. [Online]. Available: <https://situr.boyaca.gov.co/>
- [112] M. Porter, *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. Free Press, 1980.
- [113] Y. Jabareen, “Building a Conceptual Framework: Philosophy, Definitions, and Procedure,” *Int. J. Qual. Methods*, vol. 8, 11 2008.
- [114] International Organization for Standardization, *ISO 9000:2015 Quality management systems — Fundamentals and vocabulary*. ISO, 2015. [Online]. Available: <https://www.iso.org>
- [115] ——. (2015) ISO 9001:2015 Quality Management Systems - Requirements. [Online]. Available: <https://www.iso.org>
- [116] ——. (2007) Traceability in the feed and food chain – General principles and basic requirements for system design and implementation. ISO 22005:2007. [Online]. Available: <https://www.iso.org>
- [117] Food Standards Agency. (2021) TraceabilityCourse. [Online]. Available: <https://traceabilitytraining.food.gov.uk/index.html>
- [118] International Civil Aviation Organization, “Guidelines on Passenger Name Record (PNR) Data,” IATA, techreport, 2010, accessed on 17 April 2019. [Online]. Available: https://www.iata.org/contentassets/18a5fdb2dc144d619a8c10dc1472ae80/new_doc_9944_1st_edition_pnr.pdf
- [119] Food Standards Agency, “Food Traceability, Withdrawals and Recalls Within the UK Food Industry - Quick Reference Guide,” FSA, techreport, 2019, accessed on 17 April 2019. [Online]. Available: <https://www.food.gov.uk/sites/default/files/media/document/food-traceability-guide.pdf>
- [120] International Trade Centre, “Traceability in Food and Agricultural Products,” ITC, Tech. Rep. 91, 2015, accessed on 17 April 2019. [Online]. Available: <https://www.intracen.org/uploadedFiles/intracenorg/Content/Exporters/>

- [Exporting_Better/Quality_Management/Redesign/EQMBulletin91-2015_Traceability_FINAL14Oct15_web.pdf](#)
- [121] B. I. Sighencea, R. I. Stanciu, and C. D. Căleanu, “A Review of Deep Learning-Based Methods for Pedestrian Trajectory Prediction,” *Sensors*, vol. 21, no. 22, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/22/7543>
- [122] S. Wang, Z. Bao, J. S. Culpepper, and G. Cong, “A Survey on Trajectory Data Management, Analytics, and Learning,” *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1 – 36, 2021.
- [123] N. J. van Eck and L. Waltman, “Software survey: VOSviewer, a computer program for bibliometric mapping,” *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010. [Online]. Available: <https://doi.org/10.1007/s11192-009-0146-3>
- [124] H. Padrón-Ávila and R. Hernández-Martín, “How can researchers track tourists? a bibliometric content analysis of tourist tracking techniques,” *European Journal of Tourism Research*, vol. 26, pp. 2601–2601, 2020.
- [125] T. Thimm and R. Seepold, “Past, present and future of tourist tracking,” *Journal of Tourism Futures*, 03 2016.
- [126] W. Zheng, M. Li, Z. Lin, and Y. Zhang, “Leveraging tourist trajectory data for effective destination planning and management: A new heuristic approach,” *Tourism Management*, vol. 89, p. 104437, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0261517721001564>
- [127] S. Park, Y. Yuan, and Y. Choe, “Application of graph theory to mining the similarity of travel trajectories,” *Tourism Management*, vol. 87, p. 104391, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0261517721001102>
- [128] C.-P. Chu and Y.-H. Chou, “Using cellular data to analyze the tourists’ trajectories for tourism destination attributes: A case study in Hualien, Taiwan,” *Journal of Transport Geography*, vol. 96, p. 103178, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0966692321002313>
- [129] J. Chen, S. Becken, and B. Stantic, “Using Weibo to track global mobility of Chinese visitors,” *Annals of Tourism Research*, vol. 89, p. 103078, 10 2020.
- [130] S. Mikhailov and A. Kashevnik, “Car Tourist Trajectory Prediction Based on Bidirectional LSTM Neural Network,” *Electronics*, vol. 10, no. 12, 2021. [Online]. Available: <https://www.mdpi.com/2079-9292/10/12/1390>

- [131] S. Park, Y. Xu, L. Jiang, Z. Chen, and S. Huang, "Spatial structures of tourism destinations: A trajectory data mining approach leveraging mobile big data," *Annals of Tourism Research*, vol. 84, p. 102973, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0160738320301171>
- [132] C. Cay  r  , C. Faucher, C. Sallaberry, M.-N. Bessagnet, and P. Roose, "Tools for processing digital trajectories of tourists," in *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, 2020, pp. 232–233.
- [133] R. Eccleston, A. Hardy, and S. Hyslop, "Unlocking the Potential of Tracking Technology for Co-created Tourism Planning and Development: Insights from the Tourism Tracer Tasmania Project," *Tourism Planning & Development*, vol. 17, pp. 1–14, 11 2019.
- [134] R. J. Buning and V. Lulla, "Visitor bikeshare usage: tracking visitor spatiotemporal behavior using big data," *Journal of Sustainable Tourism*, vol. 29, no. 4, pp. 711–731, 2021. [Online]. Available: <https://doi.org/10.1080/09669582.2020.1825456>
- [135] T. Sakouhi, J. Malki, and J. Akaichi, "A Mobility Data Model for Web-Based Tourists Tracking," in *Doctoral Consortium/Forum@DB&IS*, 2020.
- [136] H. Chen, Y. Fan, J. Jiang, and X. Chen, "Mobility Prediction Based on POI-Clustered Data," in *Machine Learning and Intelligent Communications*, 10 2018.
- [137] S. Jang and I. Joe, "The BDD Navigation Tracking Systems Using the Beacon," in *Lecture Notes in Electrical Engineering*, 06 2018, pp. 31–36.
- [138] N. Shoval and B. McKercher, *Implementation of Tracking Technologies for Temporal and Spatial Management of Cultural Destinations: Hong Kong as an Example*. Cham: Springer International Publishing, 2017, pp. 281–294. [Online]. Available: https://doi.org/10.1007/978-3-319-09096-2_19
- [139] S. Tiwari and S. Kaushik, "Popularity estimation of interesting locations from visitor's trajectories using fuzzy inference system," *Open Computer Science*, vol. 6, 01 2016.
- [140] R. Schuitemaker and X. Xu, "Product traceability in manufacturing: A technical review," *Procedia CIRP*, vol. 93, pp. 700–705, 01 2020.
- [141] GS1, *The GS1 Traceability Standard: What you need to know*. GS1 Global Office, 2007.

- [142] J. F. Mendoza-Moreno, L. Santamaria-Granados, A. Fraga Vázquez, and G. Ramirez-Gonzalez, “OntoTouTra: Tourist Traceability Ontology Based on Big Data Analytics,” *Applied Sciences*, vol. 11, no. 22, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/22/11061>
- [143] M. Fernández-López, A. Gomez-Perez, and N. Juristo, “Methontology: from ontological art towards ontological engineering,” *Engineering Workshop on Ontological Engineering (AAAI97)*, 03 1997.
- [144] B. Kumara, I. Paik, J. Zhang, T. H. A. Siriweera, and K. Koswatte, “Ontology-Based Workflow Generation for Intelligent Big Data Analytics,” in *Conference: IEEE International Conference on Web Services (ICWS 2015)*, 06 2015.
- [145] Booking. (2019) Booking.com home page. [Online]. Available: <https://www.booking.com/>
- [146] Expedia. (2019) Expedia.com home page. [Online]. Available: <https://www.expedia.com/>
- [147] Airbnb. (2019) Airbnb.com home page. [Online]. Available: <https://www.airbnb.com/>
- [148] TripAdvisor. (2019) TripAdvisor.com home page. [Online]. Available: <https://www.tripadvisor.com/>
- [149] MinCIT. (2019) Prestadores Registro Nacional de Turismo - Datos Abiertos. [Online]. Available: <https://www.datos.gov.co/Comercio-Industria-y-Turismo/Prestadores-Registro-Nacional-de-Turismo/npkw-6rke>
- [150] Y. Bermudez, A. Aponte, V. Zuluaga, C. Moreno, and O. Ceballos. (2019) Prototipo de publicación de datos turísticos apoyados en linked open data para el consumo de información del sector ecoturístico en el centro del Valle del Cauca. [Online]. Available: <https://bibliotecadigital.univalle.edu.co/handle/10893/14492>
- [151] Ministerio de Comercio, Industria y Turismo. (2019) Informes de Turismo. [Online]. Available: <https://www.mincit.gov.co/estudios-economicos/estadisticas-e-informes/informes-de-turismo>
- [152] M. Osorio and D. Garijo. (2019) Ontology-Based APIs (OBA). [Online]. Available: <https://oba.readthedocs.io/en/latest/>
- [153] M. Musen, “The protégé project: A look back and a look forward,” *AI Matters*, vol. 1, pp. 4–12, 06 2015.

- [154] J. Hardi. (2019) Cellfie Plugin. [Online]. Available: <https://github.com/protegeproject/cellfie-plugin>
- [155] A. Gomez-Perez, M. Fernández-López, and O. Corcho, *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web*. Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web, 01 2004.
- [156] C. Steiner and D. Albert, “Validating domain ontologies: A methodology exemplified for concept maps,” *Cogent Education*, vol. 4, 01 2017.
- [157] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, and Z. Wang, “Hermit: An owl 2 reasoner,” *J. Autom. Reason.*, vol. 53, no. 3, p. 245–269, Oct. 2014.
- [158] D. Loshin, *Big Data Analytics*, E. Ltd, Ed. Amsterdam, The Netherlands: Morgan Kaufmann, 2013.
- [159] T. Bornhorst, J. Ritchie, and L. Sheehan, “Determinants of tourism success for dmos & destinations: An empirical examination of stakeholders’ perspectives,” *Tourism Management*, vol. 31, pp. 572–589, 10 2010.
- [160] C. Emani, N. Cullot, and C. Nicolle, “Understandable Big Data: A survey,” *Computer Science Review*, vol. 17, 06 2015.
- [161] P. Ceravolo, A. Azzini, M. Angelini, T. Catarci, P. Cudre-Mauroux, E. Damiani, A. Mazak, M. Van Keulen, M. Jarrar, G. Santucci, K.-U. Sattler, M. Scannapieco, M. Wimmer, R. Wrembel, and F. Zaraket, “Big data semantics,” *Journal on Data Semantics*, vol. 7, 06 2018.
- [162] V. Lytvyn, V. Vysotska, O. Veres, O. Brodyak, and O. Oryshchyn, “Big Data analytics ontology,” *Technology audit and production reserves*, vol. 1, pp. 16–27, 12 2017.
- [163] T. R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing?” *International Journal of Human-Computer Studies*, vol. 43, no. 5, pp. 907–928, 1995. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1071581985710816>
- [164] O. O. Classification, “Birte Glimm and Ian Horrocks and Boris Motik and Giorgos Stoilos,” in *Proc. of the 9th Int. Semantic Web Conf. (ISWC 2010)*, ser. LNCS, P. F. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Z. Pan, I. Horrocks, and

- B. Glimm, Eds., vol. 6496. Shanghai, China: Springer, November 7–11 2010, pp. 225–240.
- [165] M. Poveda-Villalón, A. Gomez-Perez, and M. C. Suárez-Figueroa, “OOPS! (OntOlogy Pitfall Scanner!): An on-line tool for ontology evaluation,” *International Journal on Semantic Web and Information Systems*, vol. 10, pp. 7–34, 04 2014.
- [166] J. Bandeira, I. Bittencourt, P. Espinheira, and S. Isotani, “FOCA: A Methodology for Ontology Evaluation,” *Applied Ontology*, 12 2016.
- [167] S. Ferrari and F. Cribari-Neto, “Beta Regression for Modelling Rates and Proportions,” *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, 2004. [Online]. Available: <https://doi.org/10.1080/0266476042000214501>
- [168] C. Bezerra, F. Freitas, and F. Santana da Silva, “Evaluating Ontologies with Competency Questions,” *In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 284–285, 11 2013.
- [169] Office for National Statistics, *Measuring Tourism Locally*, S. White, Ed. ONS, 2010.
- [170] UNWTO. (2019) Country Fact Sheets - Colombia. [Online]. Available: <https://webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2020-10/colombia.pdf>
- [171] ——. (2019) Tourism seasonality across destinations. [Online]. Available: <https://www.unwto.org/seasonality>
- [172] T. Tantau, *The TikZ and PGF Packages - Manual for version 3.1.9a*. Institut für Theoretische Informatik. Universität zu Lübeck, May 2021. [Online]. Available: <https://mirrors.ucr.ac.cr/CTAN/graphics/pgf/base/doc/pgfmanual.pdf>
- [173] M. Chaves and C. Trojahn, “Towards a Multilingual Ontology for Ontology-driven Content Mining in Social Web Sites,” *In Proceedings of the ISWC 2010 Workshops*, 10 2010.
- [174] M.-A. Sicilia, *Handbook of Metadata, Semantics and Ontologies*. World Scientific: Singapore, 01 2013.
- [175] L. Santamaria-Granados, J. F. Mendoza-Moreno, A. Chantre-Astaiza, M. Muñoz-Organero, and G. Ramirez-Gonzalez, “Tourist Experiences Recommender System Based on Emotion Recognition with Wearable Data,” *Sensors*, vol. 21, no. 23, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/23/7854>

- [176] A. F. Hussein, N. Arunkumar, C. Gomes, A. K. Alzubaidi, Q. A. Habash, L. Santamaria-Granados, J. F. Mendoza-Moreno, and G. Ramirez-Gonzalez, “Focal and non-focal epilepsy localization: A review,” *IEEE Access*, vol. 6, pp. 49 306–49 324, 2018.
- [177] J. Brank, M. Grobelnik, and D. Mladenić, “A survey of ontology evaluation techniques,” -, 01 2009.
- [178] M. Zarate, G. Braun, P. Fillottrani, C. Delrieux, and M. Lewis, “BiGe-Onto: An ontology-based system for managing biodiversity and biogeography data,” *Applied ontology*, pp. 1–27, 05 2020.
- [179] Booking. (2019) Trip Terms and Conditions. [Online]. Available: <https://www.booking.com/content/terms.html>
- [180] V. Krotov and L. Silva, “Legality and Ethics of Web Scraping,” in *In Proceedings of the Twenty-fourth Americas Conference on Information Systems*, 09 2018.
- [181] D. K. Mahto and L. Singh, “A dive into Web Scraper world,” in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 689–693.

Appendix A

OntoTouTra Implementation - Supplementary Material

This document relates the SPARQL queries and their results that we apply to the OntoTouTra ontology to validate it from the conceptual point of view from two approaches: The data-oriented one and the test of the ontology through competency questions. We also describe the tools we use to perform these queries, either directly through SPARQL from an endpoint or through the REST API.

This document is supplementary material to the paper: "OntoTouTra: Tourist Traceability Ontology based on Big Data Analytics." A set of ten test cases was configured, one for each KPI or CQ chosen for the ontology domain.

A.1 OntoTouTra conceptual evaluation

To validate the efficiency of OntoTouTra, we created a set of tests to verify the conceptual model and a use case. This validation is based on two approaches: The data-driven [177], where real situations from the ontology domain are represented, and in the second approach, the ontology test [178], we answer competency questions (CQ) formulated by domain experts.

In the case of the [OntoTouTra](#) data-oriented test, we created individuals gathered through Web Scraping from an [OTA](#) (Booking.com). We collected data on the destination, accommodations, services, experiences, ratings, and reviews. We chose Colombia as a use case. Therefore we filtered the data from the [OTA](#) through this country.

Regarding the set of competence questions formulated by experts in the field of tourism, we used as sources the [UNWTO](#) and ONS, and data providers of the tourism sector in Colombia, such as the Ministry of Industry, Commerce and Tourism and platforms as

SITUR. The bank of **KPI** of the ONS [169] has four categories or boxes: Satisfaction, Economy, sustainability, and Organizational. We chose and adapted ten **KPI** from this bank closely related to **OntoTouTra**'s domain: tourism traceability. For each **KPI**, we elaborated a test case, implementing the respective **SPARQL** query. After an extensive review of sources from the tourism experts and authorities noted above, the key competency questions of the ontology were specified as follows:

- KPI01: What percentage of visitors are satisfied with the provider's services?
- KPI02: What percentage of users are satisfied with the provider's internet services?
- KPI03: Number of daily visitors.
- KPI04: Impact on the destination of the offer of accommodation companies used by visitors.
- KPI05: Impact of visits on the destination.
- KPI06: Influence of accommodation companies in the destination.
- KPI07: Arrival of foreign tourists (FTA)
- KPI08: Inbound and local tourism.
- KPI09: Seasonality patterns in the destination.
- KPI10: Portfolio of tourist experiences used.

Subsequently, to solve the question of each **KPI**, we elaborated the **SPARQL** query and executed it on an **OntoTouTra** endpoint, and the results of these consulted were compared with the data obtained from the source of the domain expert. In this way, we demonstrated the effectiveness of ontology from a conceptual point of view.

Table 6.3 depicts the test cases for each of the selected **KPI**. As a reference for comparison, local government and **UNWTO** sources were sought to contrast the expected results. The test cases were run using **SPARQL** queries whose results demonstrated the reliability of the ontology when compared with the expected results.

A.2 Test cases

A.2.1 Test case 1: What percentage of visitors are satisfied with the provider's services?

Listing A.1 [KPI-01](#) % of visitors who rate the overall visitor experience as good or excellent

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ((COUNT(?goodratings)) / (COUNT(?allratings))) AS ?percentage)
WHERE {
    {
        ?review ott:hotelReviewRating ?allratings
    }
    UNION
    {
        ?review ott:hotelReviewRating ?goodratings
        FILTER(?goodratings >= 8) }
    }
```

Result: "0.715575218258312043262349"^^xsd:decimal

Interpretation: 71.5% of the reviews were rated greater than or equal to 8.0 (Good).

A.2.2 Test case 2: What percentage of users are satisfied with the provider's internet services?

Listing A.2 KPI-02 % of customers who consider the overall impression of the WiFi service to be good or excellent

```

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ((COUNT(?score)) / (COUNT(?wifiscores))) AS ?percentage
WHERE {
    {
        ?category    rdf:type    ott:ScoreCategory          ;
                    ott:scoreCategoryDescription "Free WiFi"@en .
        ?wifiscores  ott:hasScoreCategory ?category          .
    }
    UNION
    {
        ?category    rdf:type    ott:ScoreCategory          ;
                    ott:scoreCategoryDescription "Free WiFi"@en .
        ?hotelScore  ott:hasScoreCategory ?category          ;
                    ott:score    ?score                      .
        FILTER(?score >= 8)
    }
}

```

Result: "0.535060294774452880750335"^^xsd:decimal

Interpretation: 53.5% of customers consider the WiFi service as good or excellent (≥ 8).

A.2.3 Test case 3: Number of daily visitors

Listing A.3 KPI-03 Number of day visitors (Visitors who reviewed) frame

```

PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?date (COUNT(?user) AS ?visitors)
WHERE {
    ?review rdf:type ott:HotelReview ;
            ott:hotelReviewDate ?date ;
            ott:hotelReviewUser ?user .
    FILTER(?date > "2019-01-01T00:00:00"^^xsd:dateTime)
}
GROUP BY ?date
ORDER BY ASC(?date)
LIMIT 10

```

Expert: UNWTO

COLOMBIA

NOTE: Please interpret with caution. For the full data set, including metadata and footnotes, please refer to the UNWTO Database and the Methodological Notes to the UNWTO Database, available through the UNWTO website

INBOUND TOURISM

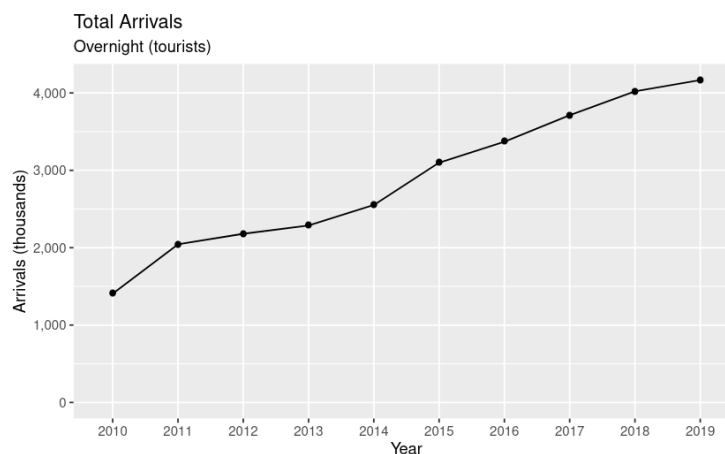


Fig. A.1 UNWTO - Country Fact Sheets: Colombia - In bound tourism

Result:

	date	visitors
1	"2019-01-02T00:00:00"^^xsd:dateTime	"2744"^^xsd:integer
2	"2019-01-03T00:00:00"^^xsd:dateTime	"2380"^^xsd:integer
3	"2019-01-04T00:00:00"^^xsd:dateTime	"1736"^^xsd:integer
4	"2019-01-05T00:00:00"^^xsd:dateTime	"2118"^^xsd:integer
5	"2019-01-06T00:00:00"^^xsd:dateTime	"2304"^^xsd:integer
6	"2019-01-07T00:00:00"^^xsd:dateTime	"2904"^^xsd:integer
7	"2019-01-08T00:00:00"^^xsd:dateTime	"3334"^^xsd:integer
8	"2019-01-09T00:00:00"^^xsd:dateTime	"2413"^^xsd:integer
9	"2019-01-10T00:00:00"^^xsd:dateTime	"2332"^^xsd:integer
10	"2019-01-11T00:00:00"^^xsd:dateTime	"1965"^^xsd:integer

Interpretation: From the expert's source, we can see 4,100,000 visitors in 2019. The execution of this query gives us a daily average of 2,423 reviews. Therefore, Booking reviews represent 21.57% of visitors to Colombia.

A.2.4 Test case 4: Impact on the destination of the offer of accommodation companies used by visitors

Listing A.4 KPI-04 Number of tourism enterprises (accommodation) per 10000 population

```

PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ((COUNT(?hotel) / xsd:integer(?population)*10000) AS ?hotels)
WHERE {
    ?hotel rdf:type ott:Hotel ;
           ott:hasCityParent ?city .
    ?city ott:hasStateParent ?state .
    ?state ott:hasCountryParent ?country .
    ?country ott:countryName ?countryName .
    ?geo gn:alternateName ?alternateName ;
        gn:population ?population .
    FILTER(?countryName = "Colombia") .
    FILTER(CONTAINS(?alternateName, ?countryName)) .
    FILTER(LANG(?alternateName) = "es") .
}
GROUP BY ?population

```

Expert: UNWTO

TOURISM INDUSTRIES

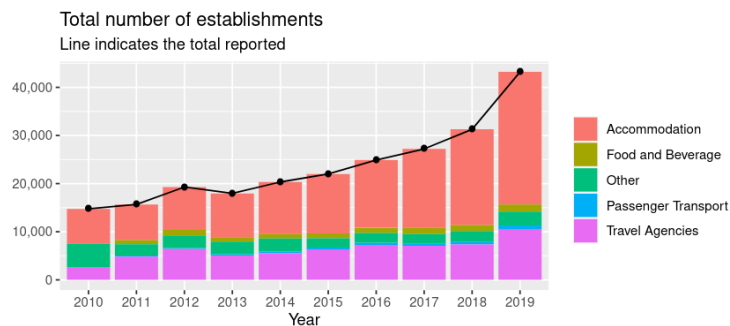


Fig. A.2 UNWTO - Country Fact Sheets: Colombia - Accommodation companies

Result: "2.32755409332593602429"^^xsd:decimal

Interpretation: From the expert's source, we can see around 28,000 accommodation establishments. Considering the population of 50 million inhabitants for Colombia in 2019,

we would have a ratio of 5.6 establishments for every 10,000 inhabitants. After executing the query, we obtained a proportion of 2.33 establishments, which means that more than half of the accommodation establishments are registered with Booking.com.

A.2.5 Test case 5: Impact of visits on the destination

Listing A.5 KPI-05 Ratio of number of reviews to local population

```

PREFIX gn:      <http://www.geonames.org/ontology#>
PREFIX wgs84:  <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX ott:    <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf:    <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd:    <http://www.w3.org/2001/XMLSchema#>
PREFIX owl:  <http://www.w3.org/2002/07/owl#>

SELECT
  ?stateName
  (MAX(COALESCE(?reviews, 0)) AS ?rev)
  (MAX(COALESCE(?statePopulation, 0)) AS ?pop)
{
  {
    SELECT ?stateName (SUM(?hotelReviewNumber) AS ?reviews)
    WHERE {
      ?hotel    ott:hotelReviewNumber ?hotelReviewNumber ;
                ott:hasCityParent      ?city                .
      ?city     ott:hasStateParent      ?state                .
      ?state    ott:stateName           ?stateName            ;
                ott:hasCountryParent   ?country              .
      ?country  ott:countryName         ?countryName         .
      FILTER(?countryName = "Colombia")
    }
    GROUP BY ?stateName
    ORDER BY ?stateName
  }
UNION
{
  SELECT ?stateName (SUM(?population) AS ?statePopulation) {
    SELECT DISTINCT ?cityName ?stateName ?population
    WHERE {
      ?hotel    ott:hasCityParent      ?city                .
      ?city     ott:cityName            ?cityName           ;
                ott:hasStateParent      ?state                .
      ?state    ott:stateName           ?stateName            ;
                ott:hasCountryParent   ?country              .
      ?country  ott:countryName         ?countryName         .
      ?geo      gn:name                 ?name                ;
                gn:population          ?geopopulation        ;
                gn:parentFeature       ?parent                .
      ?parent   gn:name                 ?parentName         .
    }
  }
}

```



```

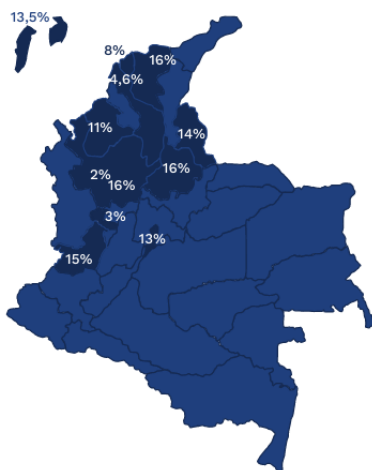
    BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
      ?name, "ú", "u", "i"),
      "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
      "á", "a", "i") AS ?acc_name)
    BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
      (?cityName, "ú", "u", "i"),
      "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
      "á", "a", "i"), "city", "", "i"), "DC", ""))
      AS ?acc_cityName)
    FILTER(CONTAINS(?acc_name, ?acc_cityName))
    BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
      ?parentName, "ú", "u", "i"),
      "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
      "á", "a", "i") AS ?acc_parentName)
    BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
      ?stateName, "ú", "u", "i"),
      "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
      "á", "a", "i") AS ?acc_stateName)
    FILTER((CONTAINS(?acc_parentName, ?acc_stateName)))
    FILTER(?countryName = "Colombia")
    BIND(xsd:integer(?geopopulation) AS ?population)
  }
}
GROUP BY ?stateName
ORDER BY ?stateName
}
}
GROUP BY ?stateName
ORDER BY ?stateName

```

Expert: MinCIT

Ministerio de Comercio, Industria y Turismo

Llegadas de pasajeros nacionales en vuelos regulares por principales aeropuertos



Aeropuerto	2018	2019	% Var
Bogotá - Eldorado	8.504.778	9.589.282	12,8%
Rionegro - Jose M. Córdoba	2.967.374	3.445.580	16,1%
Cartagena - Rafael Núñez	2.133.115	2.230.225	4,6%
Cali - Alfonso Bonilla Aragón	1.797.887	2.073.893	15,4%
Barranquilla-E. Cortissoz	1.067.791	1.157.785	8,4%
Santa Marta - Simón Bolívar	977.562	1.134.924	16,1%
San Andres - Gustavo Rojas Pinilla	944.013	1.071.036	13,5%
Bucaramanga - Palonegro	740.434	864.320	16,7%
Pereira - Matecañas	746.993	773.019	3,5%
Medellín - Olaya Herrera	520.261	530.615	2,0%
Montería - Los Garzones	456.723	505.604	10,7%
Otros	2.492.099	2.884.664	15,8%
Total General	23.349.030	26.260.947	12,5%

Fuente: Aeronáutica Civil: Boletín origen-destino, diciembre 2019. Cálculos OEE - MinCIT.

Fig. A.3 MinCIT - Colombia - Local Arrivals

Result:

stateName	reviews	population
Amazonas	5,305	44,815
Antioquia	146,842	3,830,053
Arauca	330	75,557
Atlántico	33,620	1,941,838
Bogotá	174,776	6,840,116
Bolívar	164,167	1,223,076
Boyacá	39,191	864,913
Caldas	11,254	720,124
Caquetá	419	143,871
Casanare	2,313	239,953
Cauca	5,414	448,882
Cesar	5,206	503,654
Choco	3,799	151,909
Cundinamarca	34,142	1,999,812
Córdoba	7,790	836,259

Guainía	102	17,866
Guaviare	146	53,994
Huila	14,366	722,757
La Guajira	21,765	515,117
Magdalena	98,522	656,825
Meta	9,409	598,295
Nariño	7,928	701,453
Norte de Santander	9,463	839,131
Putumayo	744	214,182
Quindío	44,912	208,314
Risaralda	20,042	866,643
San Andrés y Providencia	48,310	65,627
Santander	39,586	1,614,902
Sucre	8,780	483,695
Tolima	16,370	850,170
Valle del Cauca	44,140	3,774,893
Vaupés	26	28,382
Vichada	0	28,718

Interpretation: according to the expert, the three airports with the highest national passengers are Bogotá, Rionegro (Antioquia), and Cartagena (Bolívar). This Top-3 coincides with the query of OntoTouTra, Bogotá 174,776 reviews, Bolívar 164,167 reviews, and Antioquia 146,842 reviews.

A.2.6 Test case 6: Influence of accommodation companies in the destination

Listing A.6 KPI-06 Population rate with hotel influence

```

PREFIX gn: <http://www.geonames.org/ontology#>
PREFIX wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?stateName (?citiesPop / ?statePop * 100 AS ?populationRate){
  SELECT
    ?stateName
    (SUM(?cityPopulation) AS ?citiesPop)
    (MAX(?statePopulation) as ?statePop) {
      SELECT DISTINCT
        ?cityName ?stateName ?cityPopulation ?statePopulation
      WHERE {
        ?hotel ott:hasCityParent ?city .
        ?city ott:cityName ?cityName ;
            ott:hasStateParent ?state .
        ?state ott:stateName ?stateName ;
            ott:hasCountryParent ?country .
        ?country ott:countryName ?countryName .
        ?geo gn:name ?name ;
            gn:population ?population ;
            gn:parentFeature ?parent .
        ?parent gn:name ?parentName ;
            gn:population ?parentPopulation .
      BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
        ?name, "ú", "u", "i"),
        "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
        "á", "a", "i")
        AS ?acc_name)
      BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
        (?cityName, "ú", "u", "i"),
        "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
        "á", "a", "i"), "city", "", "i"), "DC", ""))
        AS ?acc_cityName)
      FILTER(CONTAINS(?acc_name, ?acc_cityName))
      BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
        ?parentName, "ú", "u", "i"),

```

```

        "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
        "á", "a", "i") AS ?acc_parentName)
    BIND(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(
        ?stateName, "ú", "u", "i"),
        "ó", "o", "i"), "í", "i", "i"), "é", "e", "i"),
        "á", "a", "i") AS ?acc_stateName)
    FILTER((CONTAINS(?acc_parentName, ?acc_stateName)))
    FILTER(?countryName = "Colombia")
    BIND(xsd:integer(?population) AS ?cityPopulation)
    BIND(xsd:integer(?parentPopulation) AS ?statePopulation)
    }
}
GROUP BY ?stateName
ORDER BY ?stateName
}
    
```

Expert: MinCIT

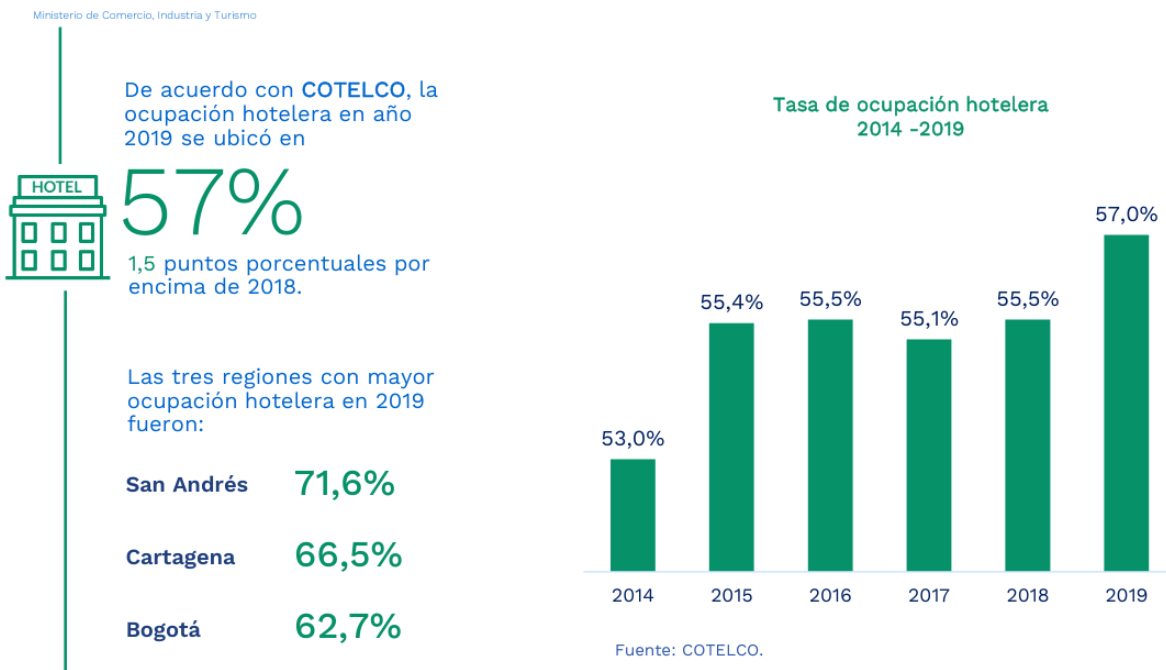


Fig. A.4 MinCIT - Colombia - Local Tourism Industry

Result:

stateName	populationRate
Amazonas	66.1710421403892153678056

Antioquia	61.8749609119911058708046
Arauca	32.5511162426007461722055
Atlántico	89.6444208081043101235553
Bogotá	100.0
Bolívar	59.688943786074063373714
Boyacá	73.3115538699174945491595
Caldas	75.4385077523380886512377
Caquetá	34.2275364766841843568375
Casanare	85.1289812529413955504092
Cauca	35.3746482291871070037362
Cesar	55.758409085122094059532
Choco	33.4579212827346210602824
Cundinamarca	77.2324395890302877519249
Córdoba	56.968627229246101139769
Guainía	50.712460970763553789384
Guaviare	56.5080428252974851126623
Huila	71.4597723196541884759812
La Guajira	78.6314052011884238711807
Magdalena	57.1193399175766598806696
Meta	85.0181570237803383182152
Nariño	45.4911164780317985727219
Norte de Santander	67.455616069454772001045
Putumayo	69.0615608837527246462796
Quindío	100.0
Risaralda	96.5609258514399298502856
San Andrés y Providencia	93.0166964311024180060663
Santander	82.7718410921708110526722
Sucre	62.6539811660470719291201
Tolima	65.3637696635714714701518
Valle del Cauca	90.7115471262848663618832
Vaupés	72.2574403625346877466331
Vichada	51.399627720504009163803

Note: In the results of Listing A.6, some departments appear with values of 100. It means that all the municipalities of that department have hotel influence.

Interpretation: In the expert's data source, the three departments with the most significant hotel influence according to their population are: San Andrés, Bolivar, and Bogotá. After executing the query, we obtained similar results except for Bolivar.

A.2.7 Test case 7: Arrival of foreign tourists (FTA)

Listing A.7 KPI-07 Foreign Tourist Arrivals (FTAs) - Top 10

```

PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?countryName (COUNT(?review) AS ?visitors)
WHERE {
    ?review ott:hotelReviewID ?hotelReviewID ;
           ott:hasCountryParent ?country .
    ?country ott:countryName ?countryName .
    FILTER(?countryName != "Colombia")
}
GROUP BY ?countryName
ORDER BY DESC(?visitors)
LIMIT 10

```

Expert: MinCIT

Ministerio de Comercio, Industria y Turismo

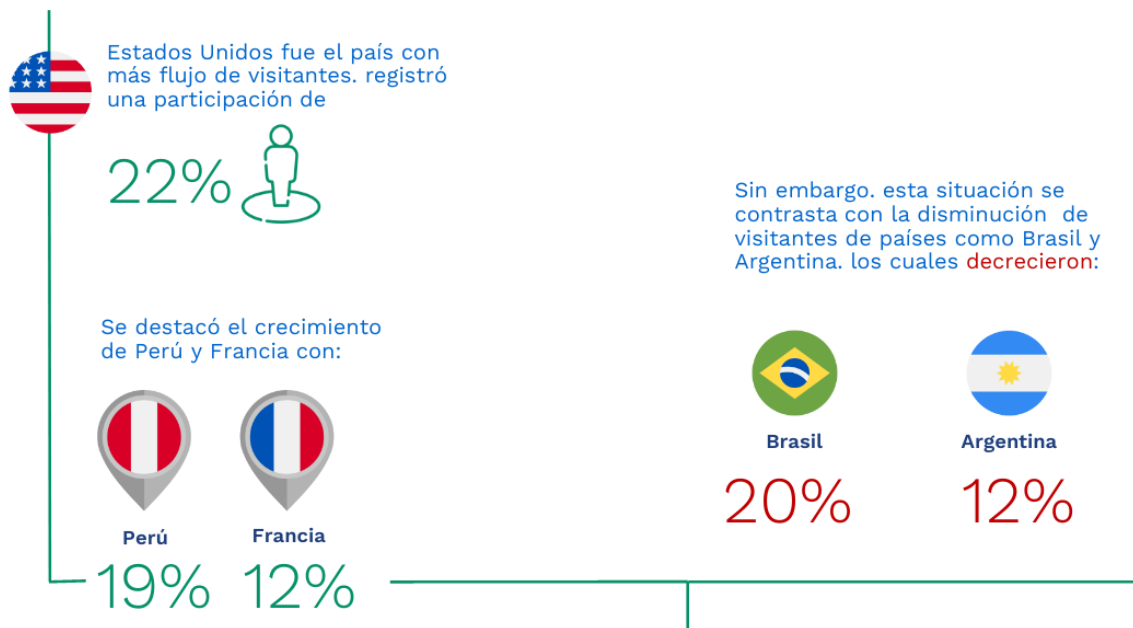


Fig. A.5 MinCIT - Colombia - FTA

Result:

countryName	visitors
USA	42,422
France	40,183
Argentina	39,753
Germany	34,225
Spain	31,420
Brazil	25,724
The Netherlands	24,447
Chile	18,878
United Kingdom	18,476
Italy	14,470

Interpretation: The provenance of foreign tourists is very similar to that reported by the expert with the query results in OntoTouTra.

A.2.8 Test case 8: Inbound and local tourism

Listing A.8 KPI-08 Inbound and domestic tourism

```

PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT
  ?stateName
  (MAX(COALESCE(?domestic, 0)) AS ?national)
  (MAX(COALESCE(?foreign, 0)) AS ?international)
{
  {
    SELECT ?stateName (COUNT(?review) AS ?domestic)
    WHERE {
      ?review ott:hotelReviewID ?hotelReviewID ;
              ott:hasHotel ?hotel ;
              ott:hasCountryParent ?country .
      ?hotel ott:hasCityParent ?city .
      ?city ott:hasStateParent ?state .
      ?state ott:stateName ?stateName .
      ?country ott:countryName ?countryName .
      FILTER(?countryName = "Colombia")
    }
    GROUP BY ?stateName
    ORDER BY ?stateName
  }
  UNION
  {
    SELECT ?stateName (COUNT(?review) AS ?foreign)
    WHERE {
      ?review ott:hotelReviewID ?hotelReviewID ;
              ott:hasHotel ?hotel ;
              ott:hasCountryParent ?country .
      ?hotel ott:hasCityParent ?city .
      ?city ott:hasStateParent ?state .
      ?state ott:stateName ?stateName .
      ?country ott:countryName ?countryName .
      FILTER(?countryName != "Colombia")
    }
    GROUP BY ?stateName
    ORDER BY ?stateName
  }
}

```

```

}
GROUP BY ?stateName
ORDER BY ?stateName

```

Expert: UNWTO See Figure A.1.

Result:

stateName	national	international
Amazonas	2,137	3,156
Antioquia	78,038	68,277
Arauca	287	42
Atlántico	25,317	8,213
Bogotá	83,918	97,460
Bolívar	62,346	101,641
Boyacá	31,753	7,445
Caldas	7,951	3,294
Caquetá	362	55
Casanare	2,145	165
Cauca	3,003	2,408
Cesar	4,740	466
Choco	2,043	1,748
Cundinamarca	30,299	4,529
Córdoba	7,108	670
Guainía	87	14
Guaviare	104	42
Huila	8,824	5,514
La Guajira	10,843	10,892
Magdalena	45,086	50,901
Meta	8,712	696
Nariño	4,216	3,703
Norte de Santander	4,493	5,078
Putumayo	472	270
Quindío	24,645	23,407
Risaralda	15,842	5,044
San Andrés y Providencia	20,610	27,601
Santander	29,755	9,792

Sucre	7,525	2,280
Tolima	15,709	1,222
Valle del Cauca	30,586	13,313
Vaupés	22	4
Total	468,978	459,322

Interpretation: According to the expert's data, inbound tourism was around 4,100,000 visitors for 2019. When executing our query, we obtained 459,322 visits, which is equivalent to 11

A.2.9 Test case 9: Seasonality patterns in the destination

Listing A.9 KPI-09 Seasonality Patterns

```

PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?month (COUNT(?review) AS ?visitors) {
    ?review ott:hotelReviewID ?reviewID ;
    ott:hotelReviewDate ?reviewDate .
    FILTER(year(?reviewDate) = 2019)
}
GROUP BY (month(?reviewDate) AS ?month)
ORDER BY ?month

```

Expert: UNWTO

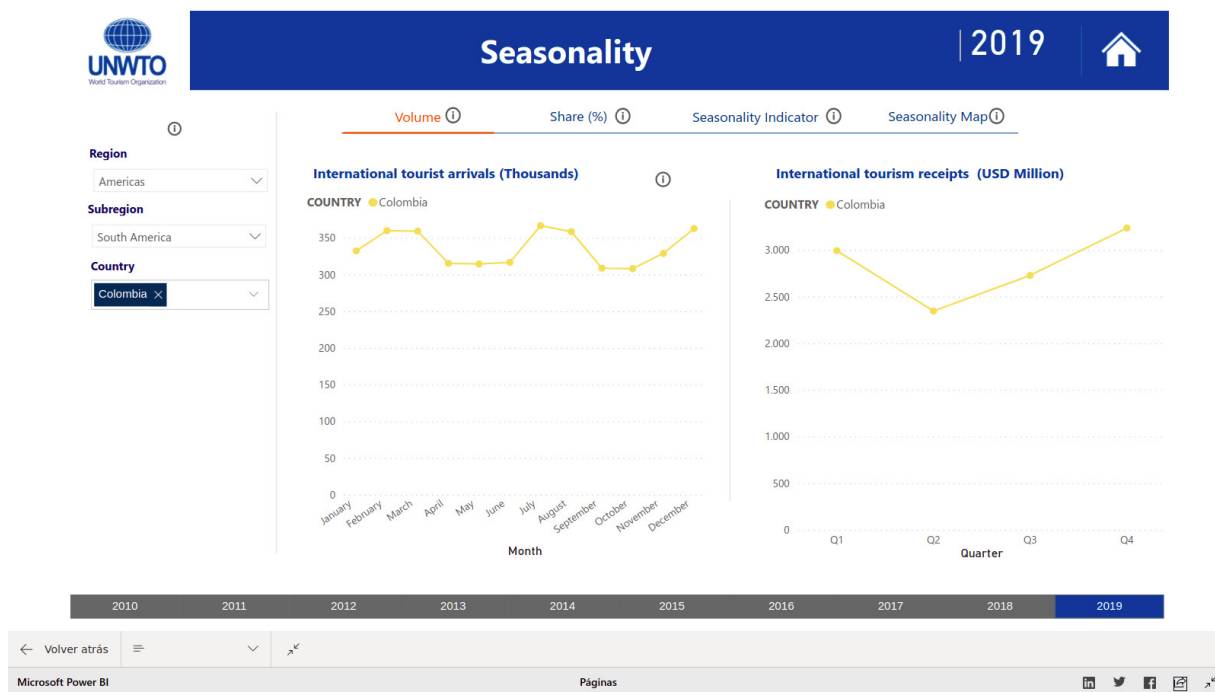


Fig. A.6 UNWTO - Colombia - Tourist Seasonality

Result:

month	reviews
01	59,177
02	41,501
03	47,009
04	46,516
05	37,277
06	41,637
07	54,123
08	60,274
09	48,215
10	47,720
11	49,521
12	41,421

Interpretation - The expert's data source matches the results of the seasonality query. Two peaks are observed per year, one between January and March and the other between July and August.

A.2.10 Test case 10: Portfolio of tourist experiences used

Listing A.10 KPI-10 Tourist experiences - Top 10

```

PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

SELECT ?activity (COUNT(?hotel) AS ?hotels) {
    ?hotel ott:hasService ?service .
    ?service ott:hasServiceCategory ?category ;
            ott:serviceName ?activity .
    ?category ott:serviceCategoryName ?categoryName .
    FILTER(STR(?categoryName) = "Activities")
    FILTER(LANG(?categoryName) = "en")
}
GROUP BY ?activity

```

Result:

activity	hotels
Beach	1,621
Walking tours	1,267
Game room	1,098
Bike tours Additional charge	1,029
Tour or class about local culture	966
Bicycle rental (for a fee)	941
Cycling Outside the accommodation	756
Trekking	752
Walking tours	698
Hiking Outside the accommodation	683

Interpretation: regarding tourist experiences, we did not find an official source from Colombia. However, our query in OntoTouTra highlighted experiences such as beach tourism, tours, and game rooms as the three most offered experiences by tourist providers in Colombia.

A.3 Big Data Analytics Lifecycle for building TTS ontology

We adapt the Big Data life cycle's Erl [4] methodology to construct our ontology in this type of environment. Below we will show some source code listings that implement the essential phases of this life cycle to illustrate this implementation with an actual use case.

A.3.1 Definition of the ontology purpose

Listing A.11 OntoTouTra preamble

```
<?xml version="1.0"?>
<rdf:RDF xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:ontotoutra="http://tourdata.org/ontotoutra/ontotoutra.owl#">
<owl:Ontology rdf:about="http://tourdata.org/ontotoutra/ontotoutra.owl">
<owl:versionIRI rdf:resource="http://tourdata.org/ontotoutra/ontotoutra.owl/1.0.0"/>
  <rdfs:comment xml:lang="en">
    tourist traceability is the analysis of the set of actions ,
    procedures , and technical measures that allows us to identify
    and record the time-space relationship of the touring ,
    from the beginning to the end of the chain of tourist products .
  </rdfs:comment>
</owl:Ontology>
```

A.3.2 Data Validation & Cleansing

In Listing A.12, we see the validation and cleaning of the general data of a hotel, in this case, the integer or float data type for the hotel or destination: id, score, and the number of reviews made.

Listing A.12 Data type validation for general destination or hotel fields

```
hotelsDF['hotelID'] = hotelsDF['hotelID'].apply(lambda x : int(x))
hotelsDF['reviewScore'] = hotelsDF['reviewScore'].apply(
  lambda x: 0 if x is None else float(x.replace(',','.'))
)
hotelsDF['reviewNumber'] = hotelsDF['reviewNumber'].apply(
  lambda x: 0 if x is None else int(x.split()[0].replace('.', ''))
)
```



```
)
```

A.3.3 Data Aggregation & Representation

In Listing A.13, we see an example of unification of datasets of reviews of some locations, using Apache Spark, which is an open-source cluster computing framework widely used in Big Data environments. We want to detect the 20 most used words in these reviews and how often these were used through clustering. The results are seen in Listing A.14.

Listing A.13 Word counts of reviews using Apache Spark

```
from pyspark import SparkContext, SparkConf

if __name__ == "__main__":
    conf = SparkConf().setAppName("word count").setMaster("local[3]")
    sc = SparkContext(conf = conf)

    lines = sc.textFile("locationOnlyReviews.txt")
    words = lines.flatMap(lambda line: line.split(" "))
    wordCounts = words.countByValue()
    result = sorted(
        wordCounts.items(), key=lambda x:x[1], reverse=True
    )

    index = 0
    for word, count in result:
        if len(word) > 0:
            print("{} : {}".format(word, count))
            index += 1
        if index >= 20:
            break
```

Listing A.14 Results of word counts of reviews using Apache Spark

```
$ spark-submit WordCount.py
the      : 2149
and      : 1913
to       : 1684
a        : 1314
is       : 1120
of       : 871
in       : 646
for      : 505
with    : 451
```

```

The      : 445
are      : 428
you      : 402
very     : 365
but      : 352
was      : 344
I        : 334
place    : 307
great    : 281
nice   : 280
town     : 279

```

A.3.4 Data analysis

In Listings A.15 and A.16, we can see the Python code snippet to convert the data from Web Scraping into triples of the [OntoTouTra](#) ontology (subject, predicate, and object). In the first four lines of the result, we can see the destination data from Web Scraping and stored in a dataset. Next, the snippet code displays the [RDF](#) representation of the destination data, as the destination belongs to a state. Through a [SPARQL](#) query, we obtained the state's name to establish the "hasStateParent" relationship. Finally, the execution of this code displays the triples in turtle format of the same [RDF](#) listing shown. The example only shows the data for one tourist destination in the dataset.

Listing A.15 Generating ontology triples with RDFLib

```

import rdflib
from rdflib.namespace import FOAF, DCTERMS, XSD, RDF, SDO
from rdflib import URIRef, BNode, Literal, Namespace

g = rdflib.Graph()
onto_filename = os.path.join(path, 'ontotoutra.owl')

format_ = rdflib.util.guess_format(onto_filename)
g.parse(onto_filename, format=format_)

qres = g.query( '''
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX my: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?stateID ?stateName
WHERE {
                ?state my:stateID ?stateID ;

```

```

        my:stateName ?stateName .
    }
'''

states = {}
for stateID, stateName in qres:
    states[stateID.value] = stateName

ott = Namespace('http://tourdata.org/ontotoutra/ontotoutra.owl#')
h = rdflib.Graph()
h.bind('ott', ott)

for index, row in cities_df.iterrows():
    cityID = row['cityID']
    cityName = row['cityName']
    stateID = row['stateID']
    stateName = ott + states[stateID]

    city = ott[cityName.replace(" ", "")]
    h.add((city, RDF.type, ott.City))
    h.add((city, ott.hasStateParent, Literal(stateName)))
    h.add((city, ott.cityID, Literal(cityID, datatype=XSD.integer)))
    h.add((city, ott.cityName, Literal(cityName)))
    h.add((city, ott.stateID, Literal(stateID, datatype=XSD.integer)))

print(row)
print(h.serialize(format='xml').decode('u8'))
print(h.serialize(format='ttl').decode('u8'))

```

Listing A.16 Results of generating ontology triples with RDFLib

```

cityID          1
cityName       Leticia
stateID         5131
Name: 0, dtype: object

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
xmlns:ott="http://tourdata.org/ontotoutra/ontotoutra.owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:about="http://tourdata.org/ontotoutra
    /ontotoutra.owl#Leticia">
    <ott:cityName>Leticia</ott:cityName>
    <rdf:type rdf:resource="http://tourdata.org

```

```

    /ontotoutra/ontotoutra.owl#City "/>
<ott:cityID rdf:datatype="http://www.w3.org/2001
/XMLSchema#integer">1</ott:cityID>
<ott:hasStateParent>
    http://tourdata.org/ontotoutra/ontotoutra.owl#Amazonas
</ott:hasStateParent>
<ott:stateID
    rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">5131
</ott:stateID>
</rdf:Description>
</rdf:RDF>

@prefix ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

ott:Leticia a ott:City ;
ott:cityID 1 ;
ott:cityName "Leticia" ;
ott:hasStateParent "http://tourdata.org/ontotoutra
/ontotoutra.owl#Amazonas" ;
ott:stateID 5131 .

```

A.3.5 Screenshots of SPARQL queries

Figures A.7, A.8, A.9, and A.10 show the execution of the SPARQL queries on different endpoints: Apache Fuseki, Apache Jena, Protégé, and Open Link Virtuoso. It demonstrates the interoperability of the ontology.

The screenshot displays the Apache Fuseki SPARQL query interface. At the top, there are navigation links for 'query', 'upload files', 'edit', and 'info'. Below this is the 'SPARQL query' section, which includes a prompt to enter a query and a section for 'EXAMPLE QUERIES' with buttons for 'Selection of triples' and 'Selection of classes'. The 'PREFIXES' section shows buttons for 'rdf', 'rdfs', 'owl', 'xsd', and a refresh icon. The 'SPARQL ENDPOINT' is set to '/ds/query', and the 'CONTENT TYPE (SELECT)' is 'JSON' and 'CONTENT TYPE (GRAPH)' is 'Turtle'. The main area contains a SPARQL query:

```

1 PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
2 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
4 PREFIX owl: <http://www.w3.org/2002/07/owl#>
5
6 SELECT ?countryName (COUNT(?review) AS ?visitors)
7 WHERE {
8   ?review ott:hotelReviewID ?hotelReviewID ;
9           ott:hasCountryParent ?country ;
10          ?country ott:countryName ?countryName .
11  FILTER(?countryName != "Colombia")
12 }
13 GROUP BY ?countryName
14 ORDER BY DESC(?visitors)
15 LIMIT 10

```

Below the query is the 'QUERY RESULTS' section, which has buttons for 'Table' and 'Raw Response'. It shows 'Showing 1 to 10 of 10 entries' and a search box. The results are displayed in a table with columns 'countryName' and 'visitors':

	countryName	visitors
1	"Estados"	"42422"^^xsd:integer
2	"Francia"	"40183"^^xsd:integer
3	"Argentina"	"39753"^^xsd:integer

Fig. A.7 SPARQL query in OntoTouTra using Apache Fuseki

```
(base) [jf@fedora 06 ontology]$ cat q1.rq
PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT ?countryName ?p ?o
WHERE {
  ?country rdf:type      ott:Country ;
           ott:countryName ?countryName .
  ?country ?p           ?o .
  FILTER (?countryName = 'France')
}
(base) [jf@fedora 06 ontology]$ sparql --data=ontotoutra.owl --query=q1.rq

-----
| countryName | p                | o                |
-----
| "France"    | ott:alpha2Code   | "FR"             |
| "France"    | ott:alpha3Code   | "FRA"            |
| "France"    | rdf: type        | ott:Country      |
| "France"    | rdf: type        | <http://www.w3.org/2002/07/owl#NamedIndividual> |
| "France"    | ott:countryName  | "France"         |
| "France"    | ott:countryID    | 250              |
-----
```

Fig. A.8 SPARQL query in [OntoTouTra](#) using Apache Jena

The screenshot shows the OntoTouTra web application interface. The browser address bar displays the URL: `http://tourdata.org/ontotoutra/ontotoutra.owl/1.0.0`. The application has a menu bar with options: File, Edit, View, Reasoner, Tools, Refactor, Window, Ontop, Mastro. Below the menu bar, there are navigation buttons for '<' and '>', and a breadcrumb trail: 'ontotoutra (http://tourdata.org/ontotoutra/ontotoutra.owl/1.0.0)'. The main content area has tabs for 'Active ontology', 'Entities', 'Individuals by class', and 'DL Query'. The 'DL Query' tab is active, showing a 'SPARQL query' input field. The query text is:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

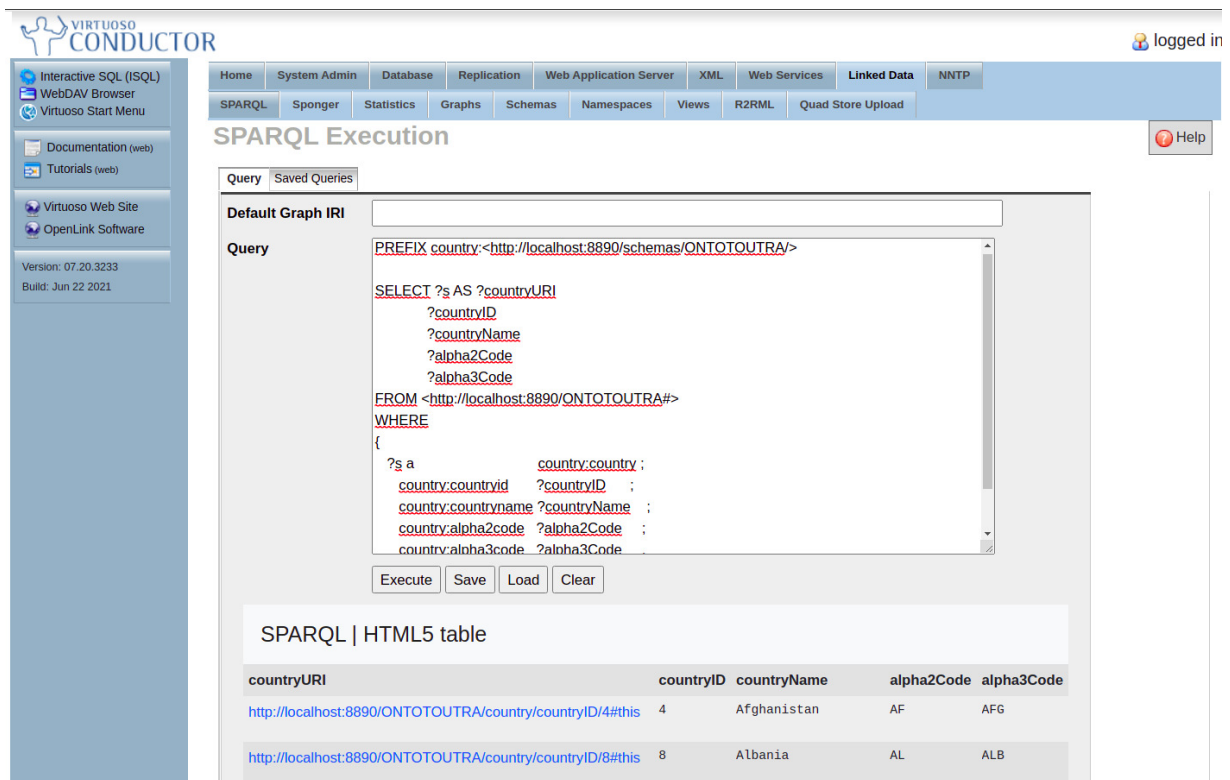
SELECT ?subject ?object
WHERE {
  ?subject rdfs:subClassOf ?object
}
```

Below the query, there is a table with two columns: 'subject' and 'object'. The table contains the following data:

subject	object
Adventure	Experience
Academic	Experience
Ecotourism	Nature
Tradition	Cultural
Ride	Adventure
Abseiling	Adventure
Accommodation	Provider

At the bottom of the query area, there is an 'Execute' button. Below the table, there are tabs for 'Ontology imports', 'Ontology Prefixes', and 'General class axioms'.

Fig. A.9 SPARQL query in OntoTouTra using Protégé



SPARQL Execution

Query

```
PREFIX country:<http://localhost:8890/schemas/ONTOTOUTRA/>

SELECT ?s AS ?countryURI
       ?countryID
       ?countryName
       ?alpha2Code
       ?alpha3Code
FROM <http://localhost:8890/ONTOTOUTRA#>
WHERE
{
  ?s a          country:country ;
    country:countryid ?countryID ;
    country:countryname ?countryName ;
    country:alpha2code ?alpha2Code ;
    country:alpha3code ?alpha3Code .
}
```

Execute Save Load Clear

SPARQL | HTML5 table

countryURI	countryID	countryName	alpha2Code	alpha3Code
http://localhost:8890/ONTOTOUTRA/country/countryID/4#this	4	Afghanistan	AF	AFG
http://localhost:8890/ONTOTOUTRA/country/countryID/8#this	8	Albania	AL	ALB

Fig. A.10 SPARQL query in OntoTouTra using OpenLink Virtuoso

```
(base) [jf@fedora 10_RESTful]$ s-query --service http://localhost:3030/
ds/query 'PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>

SELECT ?cityID ?cityName ?stateName
WHERE {
  ?state ott:stateName    ?stateName ;
         ott:stateCapital ?cityName .
  FILTER(?stateName = "Boyaca")
}'
{ "head": {
  "vars": [ "cityID" , "cityName" , "stateName" ]
} ,
  "results": {
    "bindings": [
      {
        "cityName": { "type": "literal" , "value": "Tunja" } ,
        "stateName": { "type": "literal" , "value": "Boyaca" }
      }
    ]
  }
}
```

Fig. A.11 REST API in OntoTouTra using Fuseki SOH

The screenshot shows a REST API interface for the `HotelReview` endpoint. The interface is titled "HotelReview" and displays the following information:

- Method:** GET
- Endpoint:** `/hotelreviews` (List all instances of HotelReview)
- Description:** Gets a list of all instances of HotelReview (more information in <http://tourdata.org/ontotoutra/ontotoutra.owl#HotelReview>)
- Parameters:**
 - label:** string (query), Filter by label. Input: `label - Filter by label`
 - page:** integer(\$int32) (query), Page number. Input: `1`
 - per_page:** integer(\$int32) (query), Items per page. Input: `100`
- Buttons:** Execute, Clear, Cancel
- Responses:**
 - Curl:** `curl -X GET "http://localhost:8080/v0.0.1/hotelreviews?page=1&per_page=100" -H "accept: application/json"`
 - Request URL:** `http://localhost:8080/v0.0.1/hotelreviews?page=1&per_page=100`
 - Server response:**

Code	Details
200	

Fig. A.12 REST API in OntoTouTra using OBA OpenAPI

In Figures A.13 and A.14, we see the documentation generated by OntoTouTra from two different systems: Protégé and OBA, respectively.

Ontologies Classes Object Properties Data Properties Annotation Properties
Individuals Datatypes Clouds

Contents

- ontotoutra
- Classes (67)
- Object Properties (18)
- Data Properties (111)
- Annotation Properties (1)
- Datatypes (7)

OWL HTML inside

Ontologies Classes Object Properties Data Properties Annotation Properties Individuals

Class: Hotel

Annotations (1)

- rdfs:comment "Hotel"

Superclasses (1)

- Accommodation

Disjoints (10)

AccommodationType, ApartHotel, Camping, Hostel, **Hotel**, HotelScore, LodgingHouse,

Usage (15)

- hasCityParent **Domain** (Provider or **Hotel** or Attraction)
- hasService **Domain** **Hotel**
- hasHotel **Range** **Hotel**
- hasHotelScore **Range** **Hotel**
- cityID **Domain** (**Hotel** or City)
- hotelAddress **Domain** **Hotel**
- hotelDescription **Domain** **Hotel**
- hotelID **Domain** (HotelScore or HotelReview or **Hotel**)
- hotelLat **Domain** **Hotel**
- hotelLon **Domain** **Hotel**
- hotelName **Domain** **Hotel**
- hotelReviewCategoricalScore **Domain** **Hotel**
- hotelReviewNumber **Domain** **Hotel**
- hotelReviewScore **Domain** **Hotel**
- hotelURL **Domain** **Hotel**

Ontologies Classes Object Properties Data Properties Annotation Properties
Individuals Datatypes Clouds

Class: Hotel

Annotations (1)

- rdfs:comment "Hotel"

Superclasses (1)

- Accommodation

Disjoints (10)

AccommodationType, ApartHotel, Camping, Hostel, **Hotel**, HotelScore, LodgingHouse, Resort, RuralAccommodation, TouristHousing

Usage (15)

- hasCityParent **Domain** (Provider or **Hotel** or Attraction)
- hasService **Domain** **Hotel**
- hasHotel **Range** **Hotel**
- hasHotelScore **Range** **Hotel**
- cityID **Domain** (**Hotel** or City)
- hotelAddress **Domain** **Hotel**
- hotelDescription **Domain** **Hotel**
- hotelID **Domain** (HotelScore or HotelReview or **Hotel**)
- hotelLat **Domain** **Hotel**
- hotelLon **Domain** **Hotel**
- hotelName **Domain** **Hotel**
- hotelReviewCategoricalScore **Domain** **Hotel**
- hotelReviewNumber **Domain** **Hotel**

Ontologies Classes Object Properties Data Properties Annotation Properties Individuals

Class: Hotel

Annotations (1)

- rdfs:comment "Hotel"

Superclasses (1)

- Accommodation

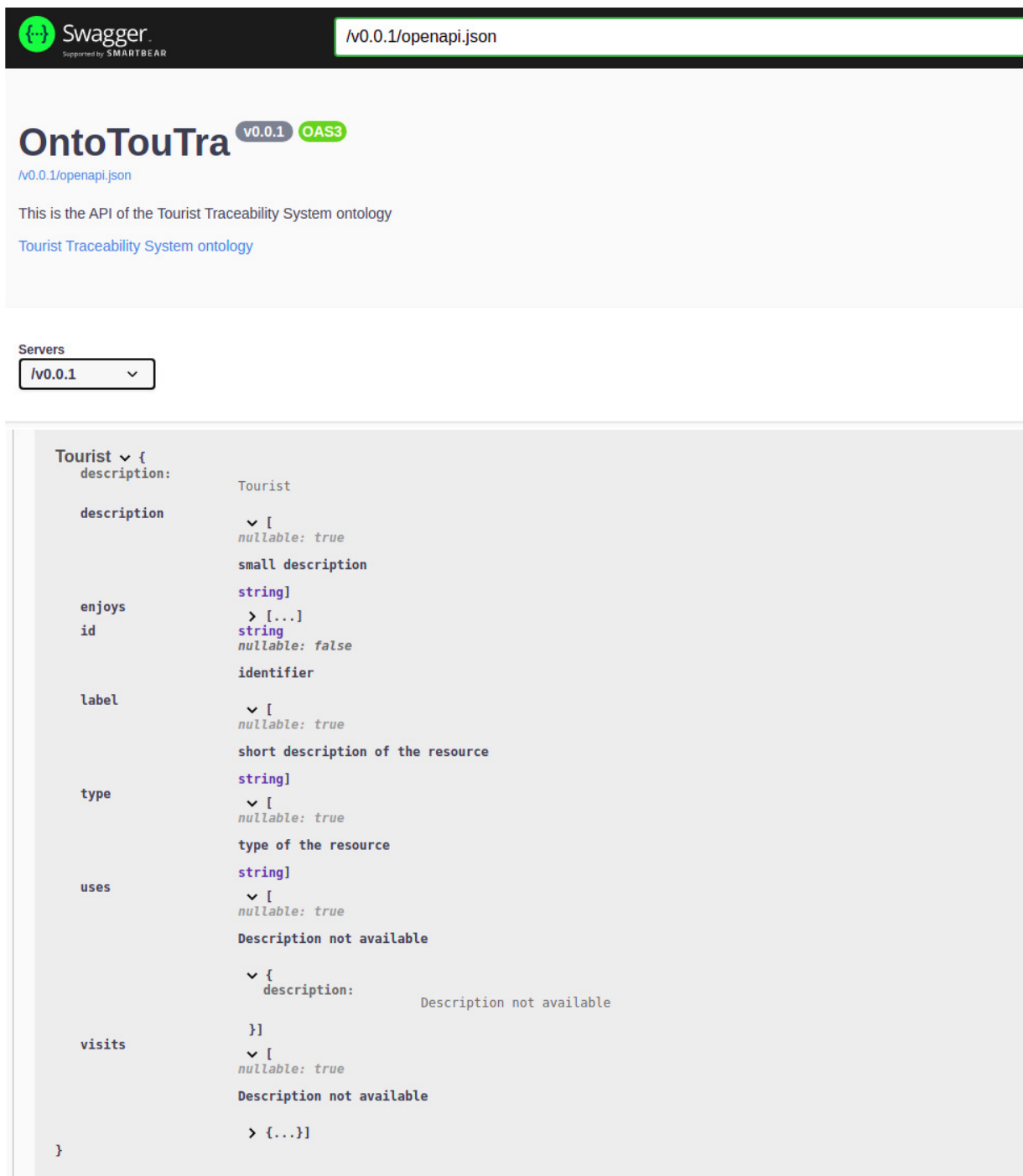
Disjoints (10)

AccommodationType, ApartHotel, Camping, Hostel, **Hotel**, HotelScore, LodgingHouse,

Usage (15)

- hasCityParent **Domain** (Provider or **Hotel** or Attraction)
- hasService **Domain** **Hotel**
- hasHotel **Range** **Hotel**
- hasHotelScore **Range** **Hotel**
- cityID **Domain** (**Hotel** or City)
- hotelAddress **Domain** **Hotel**
- hotelDescription **Domain** **Hotel**
- hotelID **Domain** (HotelScore or HotelReview or **Hotel**)
- hotelLat **Domain** **Hotel**
- hotelLon **Domain** **Hotel**
- hotelName **Domain** **Hotel**
- hotelReviewCategoricalScore **Domain** **Hotel**
- hotelReviewNumber **Domain** **Hotel**
- hotelReviewScore **Domain** **Hotel**
- hotelURL **Domain** **Hotel**

Fig. A.13 OntoTouTra documentation generated by Protégé



The image shows the Swagger UI for the OntoTouTra API. At the top, the Swagger logo is visible, along with the URL `/v0.0.1/openapi.json`. The main heading is **OntoTouTra** with version `v0.0.1` and the OAS3 specification version. Below this, it states: "This is the API of the Tourist Traceability System ontology" and provides a link to the "Tourist Traceability System ontology".

Under the "Servers" section, the selected server is `/v0.0.1`.

The main content area displays the JSON schema for the `Tourist` object:

```

Tourist {
  description: Tourist
  description:
    > [
      nullable: true
      small description
    ]
  enjoys:
    > [...]
  id: string
  identifier:
    nullable: false
  label:
    > [
      nullable: true
      short description of the resource
    ]
  type: string
  type of the resource:
    > [
      nullable: true
    ]
  uses: string
  Description not available:
    > [
      nullable: true
      Description not available
    ]
  visits:
    > [
      nullable: true
      Description not available
    ]
  > [...]
}

```

Fig. A.14 OntoTouTra documentation generate by OBA

A.4 Disclaimer

We recommend that whoever uses the software that accompanies this paper use it responsibly. To avoid legal problems, check the Web Site's rules, the data provider device, or the application installed on the user's device.

Appendix B

Beacons installation

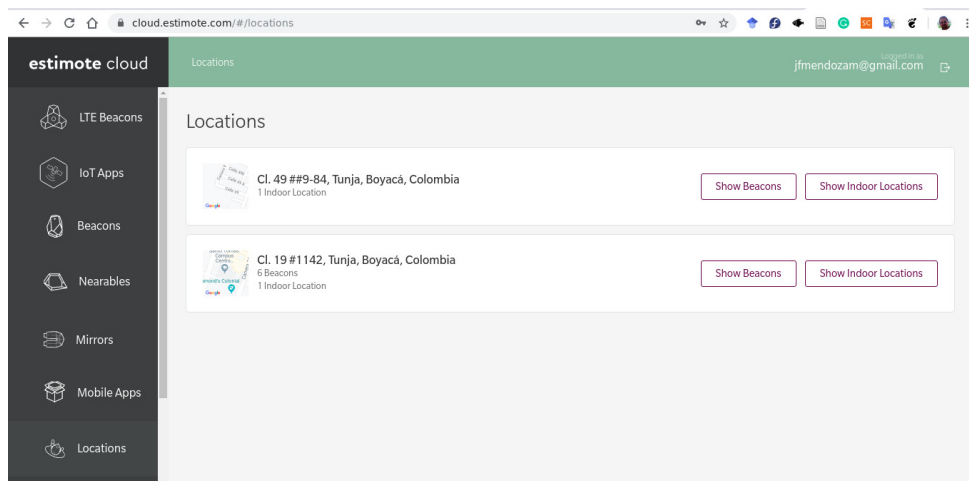


Fig. B.1 Locations of the beacons

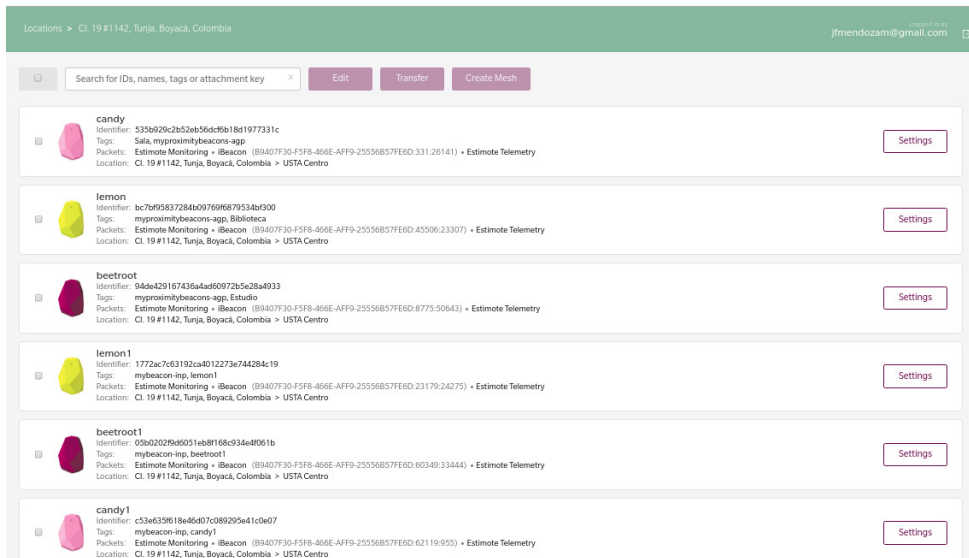


Fig. B.2 Cluster of beacons

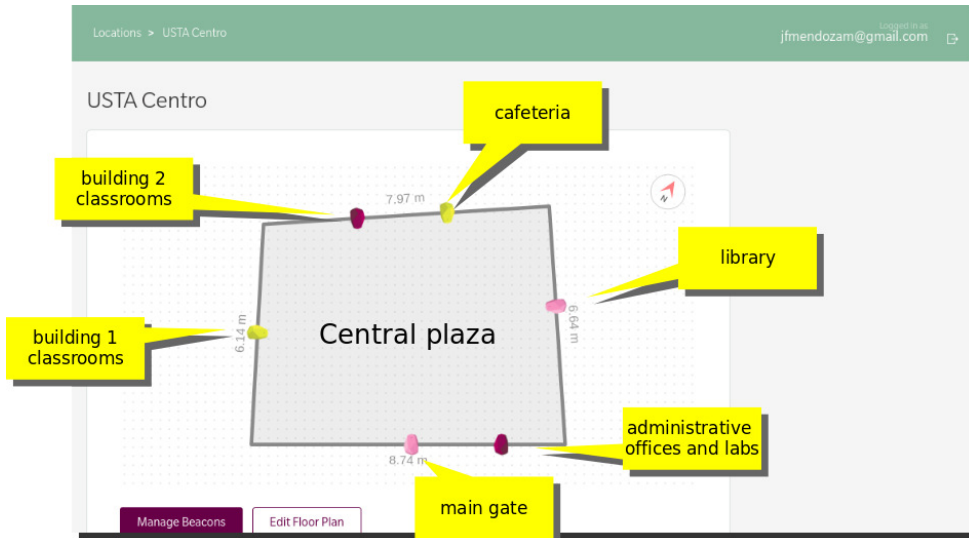


Fig. B.3 Map of beacons



Fig. B.4 Beacons in the POI

```

package com.estimate.indoorsdk_module.view

import ...

/**
 * Main view for drawing location (walls, beacons, objects, etc.) and user position on it.
 * You need to declare it in your XML layout file first and then traditionally bind it to an object.
 *
 * @author Pawel Dylag (pawel.dylag@estimate.com)
 */
@Keep
class IndoorLocationView : View {

    internal class ViewLocationPosition(val x: Double = 0.0,
                                       val y: Double = 0.0,
                                       val orientation: Double = 0.0,
                                       val isHidden: Boolean = false)

    // STYLEABLE
    private var mWallPaint = Paint(Paint.ANTI_ALIAS_FLAG)
    private var mCustomPoint = Paint(Paint.ANTI_ALIAS_FLAG)
    private var mNearbyBeaconPaint = Paint(Paint.ANTI_ALIAS_FLAG)
    private var mNearbyBeaconTextPaint = Paint(Paint.ANTI_ALIAS_FLAG)
    private var mBeaconBitmapScale = 0.3f
    private var mWallStrokeWidth = 5f
    private var mCustomPointStrokeWidth = 20f
    private var mNearbyBeaconRadius = 40f
    private var mLocationPadding = 10
    private var mPositionAnimationDuration = 800L

    // DATA OBJECTS
    private var mLocation = Location()
    private var mPosition = ViewLocationPosition(isHidden = true)
    private var mNearbyBeacons: List<BeaconWithDistance> = emptyList()
    private var mCustomPoints: List<LocationPosition> = emptyList()

    // SCREEN OBJECTS
    private val mLocationPath = Path()
    private val mNearbyBeaconsPath = Path()
    private val mBeaconColorToBitmapMapping = createColorToBitmapMapping()
    private val mPositionBitmap = createPositionBitmap()

    // CURRENT DRAW PARAMETERS
    private val mDrawParams = DrawParams()

```

Fig. B.5 Location data objects code

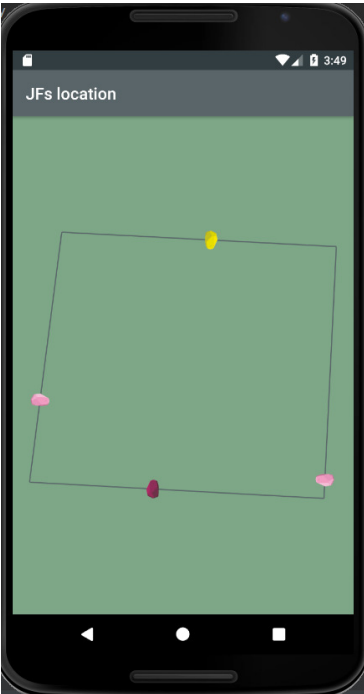


Fig. B.6 Beacon app

Appendix C

MEB App

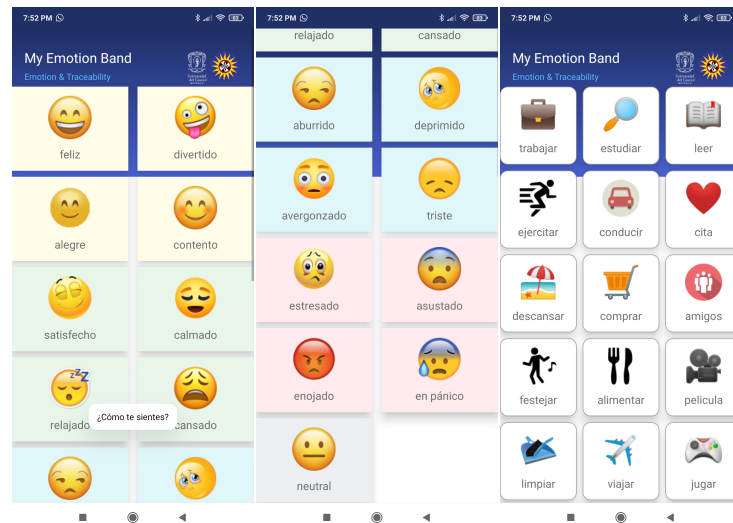


Fig. C.1 MEB app

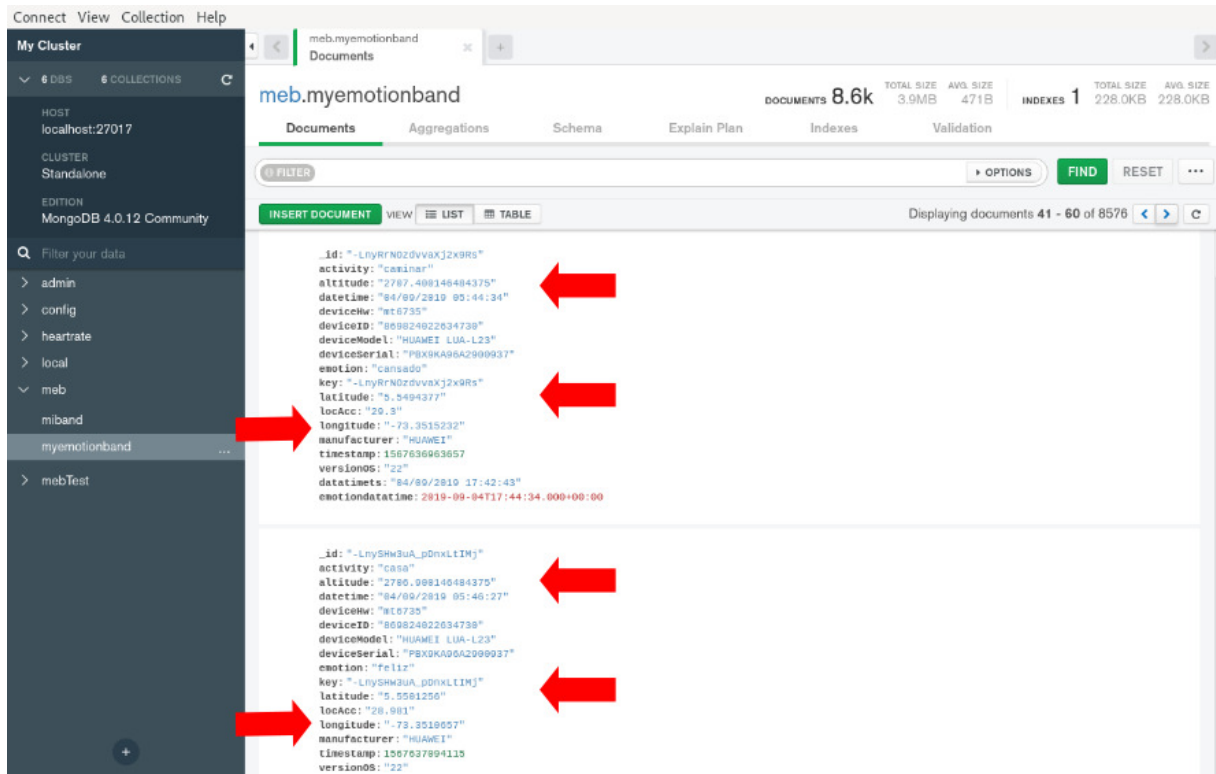


Fig. C.2 MEB dataset

Appendix D

Location intelligence component

```
import numpy as np

def haversine(lat1, lon1, lat2, lon2):
    R = 3958.76 # Earth radius in miles
    dLat = np.radians(lat2 - lat1)
    dLon = np.radians(lon2 - lon1)
    lat1 = np.radians(lat1)
    lat2 = np.radians(lat2)
    a = np.sin(dLat/2) ** 2 + np.cos(lat1) * np.cos(lat2) * np.sin(dLon / 2) ** 2
    c = 2 * np.arcsin(np.sqrt(a))
    return R * c
```

Fig. D.1 Haversine distance

```
theta1 = np.radians(-28.904)
theta2 = np.radians(28.904)
R1 = np.array([[np.cos(theta1), np.sin(theta1)], [-np.sin(theta1), np.cos(theta1)]])
R2 = np.array([[np.cos(theta2), np.sin(theta2)], [-np.sin(theta2), np.cos(theta2)]])

def manhattan_dist(lat1, lon1, lat2, lon2):
    p = np.stack([lat1, lon1], axis = 1)
    d = np.stack([lat2, lon2], axis = 1)
    pT = R1 @ p.T
    dT = R1 @ d.T

    vT = np.stack((pT[0,:], dT[1,:]))
    v = R2 @ vT
    return (haversine(p.T[0], p.T[1], v[0], v[1]) + haversine(v[0], v[1], d.T[0], d.T[1]))
```

Fig. D.2 Manhattan distance

```
In [86]: eCursor = ec.find()
for x in eCursor:
    print (x['emotiondatetime'], x['deviceID'], x['latitude'], x['longitude'], x['altitude'])

2019-09-06 09:37:41 353420084627218 5.5330424 -73.3637059 2811.7001953125
2019-09-06 09:45:41 353420084627218 5.5330433 -73.3636726 2814.400146484375
2019-09-06 09:49:11 353420084627218 5.5330433 -73.3636726 2814.400146484375
2019-09-06 09:50:18 353420084627218 5.5330433 -73.3636726 2814.400146484375
2019-09-06 10:12:18 353420084627218 5.5330433 -73.3636726 2814.400146484375
2019-09-06 10:13:41 353420084627218 5.5330433 -73.3636726 2814.400146484375
2019-09-06 16:41:55 353420084627218 5.5493933 -73.3453795 2707.60009765625
2019-09-08 12:12:27 353420084627218 5.5525198 -73.349237 0.0
2019-09-08 12:59:18 353420084627218 5.3183811 -73.3948348 0.0
2019-09-08 13:07:09 353420084627218 5.318447 -73.3947975 2093.60009765625
2019-09-08 16:54:15 353420084627218 5.317909 -73.3963679 2108.900146484375
2019-09-08 19:47:36 353420084627218 5.5345634 -73.3531942 2722.7001953125
2019-09-08 22:48:36 353420084627218 5.5345634 -73.3531942 2722.7001953125
2019-09-08 23:34:40 353420084627218 5.5345634 -73.3531942 2722.7001953125
2019-09-09 06:33:05 353420084627218 5.534592 -73.3532063 2722.7001953125
2019-09-09 06:57:44 353420084627218 5.5325224 -73.3640539 0.0
2019-09-09 09:05:46 353420084627218 5.5326006 -73.3636858 2817.10009765625
2019-09-09 13:40:27 353420084627218 5.5314905 -73.3611065 0.0
2019-09-09 14:05:08 353420084627218 5.535995 -73.3530009 0.0
```

Fig. D.3 Location dataset

```
eCursor = ec.find()
locationList = []
for locationData in eCursor:
    tmpList = [
        locationData['deviceID'],
        locationData['emotiondatetime'],
        locationData['latitude'],
        locationData['longitude'],
        locationData['altitude']
    ]
    locationList.append(tmpList)

locationDF = pd.DataFrame(locationList, columns = ['user', 'date', 'lat', 'long', 'alt'])

user = "866962043739436"
subject1Data = locationDF.loc[locationDF['user'] == user]

today = max(subject1Data['date']) - timedelta(days=1)
subject1YesterdayData = subject1Data.loc[subject1Data['date'] >= pd.to_datetime(today)]
coordDF = subject1YesterdayData[['lat', 'long']].copy()
coordDF['name'] = ['L{}'.format(i) for i in range(1, len(coordDF) + 1)]

coords = Table.from df(coordDF)
coords.select('lat', 'long', 'name')

#Marker.map_table(coords.select('lat', 'long', 'name'))
```

Fig. D.4 Location operations

Appendix E

OntoTouTra Development

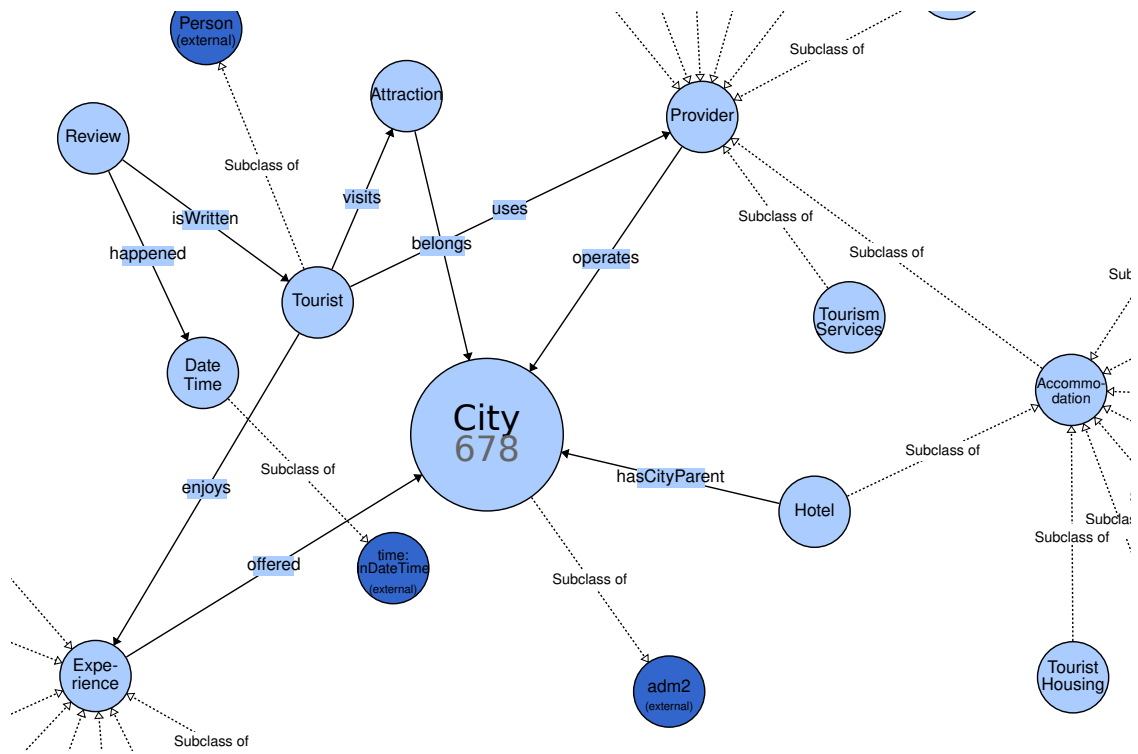


Fig. E.1 Snippet of the image of the upper levels of OntoTouTra (using WebVOWL [5]).

BookingWebScraping
<ul style="list-style-type: none"> - locationList : list - hotelList : list - language : string - location : string - directory : string - locationHeader : string - hotelHeader : string - source : string - inputFilename : string - country : string - countryName : string - web : string - cityKeyword : dict - hotelKeyword : dict - propertyTags : dict
<ul style="list-style-type: none"> + getLocation(region : string, language : string) : list + getPropertiesReview() : list + addLocation(hotelID : int, hotelName : string, hotelURL ... + getPropertiesList(locationID : int, location : string) : list + getLocationID(location : string) : int + hotel2CSV(reviews : list) + location2CSV(location : string, reviews : list) + getLocationReviews(location : string) : list + getSource() : string + getLocationsByCountry() : list + getPropertySource() + getPropertiesByCountry() : list + getCountryReviews()

Fig. E.2 Web scraping class.

```

PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX gn: <http://www.GeoNames.org/ontology#>
PREFIX wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace/>

SELECT ?city ?cityName ?stateName ?countryName
       ?geo ?geoName ?alternateName ?countryCode
       ?cityLat ?cityLon
WHERE {
  ?city    ott:cityID          ?cityID          ;
          ott:cityName       ?cityName         ;
          ott:stateID        ?stateID          ;
          ott:hasStateParent ?state            .
  ?state   ott:stateName     ?stateName        ;
          ott:hasCountryParent ?country        .
  ?country ott:countryName   ?countryName      .
  ?geo     gn:name           ?geoName          ;
          gn:alternateName   ?alternateName    ;
          gn:countryCode     ?countryCode      ;
          gn:population      ?cityPopulation  ;
          wgs84_pos:lat      ?cityLat          ;
          wgs84_pos:long     ?cityLon          .
  FILTER(?cityName = "Cartagena de Indias") .
  FILTER(CONTAINS(?alternateName, ?cityName)) .
  FILTER(LANG(?alternateName) = "es") .
}
LIMIT 1

```

Fig. E.3 Listing of Data link to GeoNames for obtaining city coordinates

```

{
  "head": {
    "vars": [
      "city", "cityName", "stateName", "countryName", "geo",
      "geoName", "alternateName", "countryCode", "cityLat", "cityLon"
    ]
  },
  "results": {
    "bindings": [
      {
        "alternateName": {
          "type": "literal",
          "value": "Municipio de Cartagena de Indias",
          "xml:lang": "es"
        },
        "city": {
          "type": "uri",
          "value": "http://tourdata.org/ontotoutra/ontotoutra.owl#CartagenaDeIndias"
        },
        "cityLat": { "type": "literal", "value": "10.50743" },
        "cityLon": { "type": "literal", "value": "-75.4543" },
        "cityName": { "type": "literal", "value": "Cartagena de Indias" },
        "countryCode": { "type": "literal", "value": "CO" },
        "countryName": { "type": "literal", "value": "Colombia" },
        "geo": { "type": "uri", "value": "https://sws.GeoNames.org/3687247/" },
        "geoName": { "type": "literal", "value": "Municipio de Cartagena de Indias" },
        "stateName": { "type": "literal", "value": "Bolivar" }
      }
    ]
  }
}

```

Fig. E.4 Results of data link to GeoNames for obtaining city coordinates

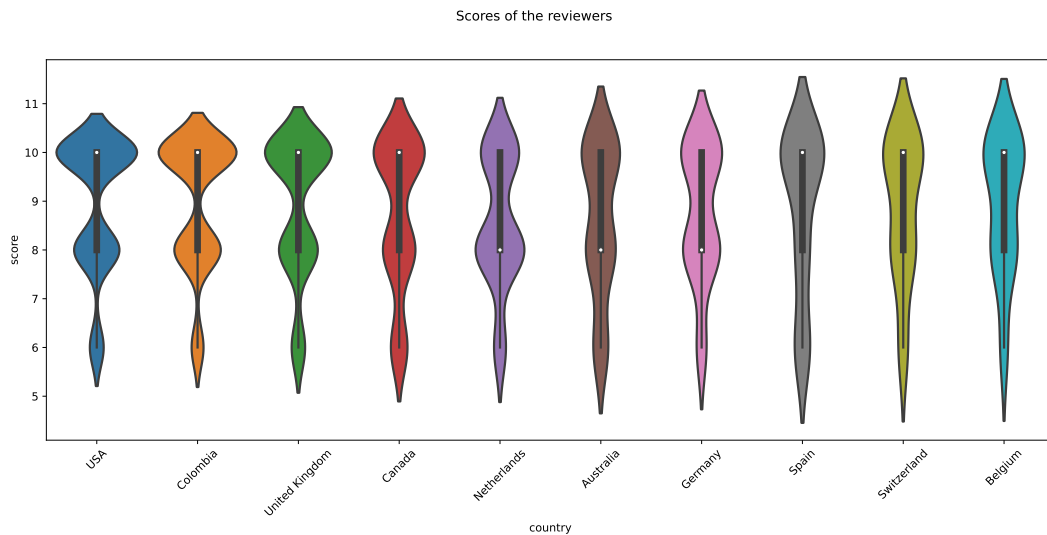


Fig. E.5 Distribution of the scores of the top 10 nationalities of reviewers of Colombia's tourist reviews dataset obtained from OntoTouTra (language: English).

```

{
  "Collections": [{
    "sheetName": "city",
    "startColumn": "A",
    "endColumn": "A",
    "startRow": "2",
    "endRow": "+",
    "comment": "",
    "rule":
      "Individual: @B*\n
Types: City\n
Facts: \n
cityID @A* (xsd:integer),\n
cityName @B*,\n
stateID @C* (xsd:integer)",
    "active": true
  }]
}

```

Fig. E.6 An example of transformation rules from the Cities spreadsheet

```

# Selenium web scraping driver
d = webdriver.Chrome(chromeDriverDir)

# Iterate states of a country
for state in states:
    hotelsDF = getHotelData(d, state, ' ')

    urlNextPage, hotelsPerPage, totalPages, hotelsNum = getNextPageURL(d)
    startInstance = len(hotelsDF) if hotelsPerPage == 0 else hotelsPerPage
    while (startInstance + 1 <= hotelsNum):
        newURL = urlNextPage + tags['offset']['tag'] + str(startInstance)
        hotelsDF = hotelsDF.append(
            getHotelData(d, state, newURL),
            ignore_index = True
        )
        startInstance += hotelsPerPage

    hotelsDF = getHotelAddress(d, hotelsDF)
    hotelsDF, hotel_services = getHotelServices(d, hotelsDF)
    hotelsDF, hotel_reviews, hotel_scores = getHotelComments(d, hotelsDF)
    hotelsDF.drop_duplicates(subset = 'hotelID', inplace = True)
    hotelsDF.reset_index(drop = True)
    servicesDF = pd.DataFrame(hotel_services)
    servicesDF['hotelID'] = hotelsDF['hotelID']
    scoresDF = pd.DataFrame(hotel_scores)
    scoresDF['hotelID'] = hotelsDF['hotelID']

# Get the keys of the dictionary list
list(dict.fromkeys([k for i in hotel_reviews[0] for k, v in i.items()]))
# Get the hotel reviews Dataframe
hreviews = []
for hotel in hotel_reviews:
    for comment in hotel:
        hreviews.append(comment)
reviewsDF = pd.DataFrame(hreviews)

```

Fig. E.7 Python code snippet about OTA web scraping


```

import pandas as pd
import rdflib

g = rdflib.Graph()
onto_filename = os.path.join(path, 'ontotoutra.owl')
format_ = rdflib.util.guess_format(onto_filename)
g.parse(onto_filename, format=format_)

qres = g.query(' '
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ott: <http://tourdata.org/ontotoutra/ontotoutra.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?hotelID ?cityName ?stateName ?score ?lat ?lon
WHERE {
    ?hotel ott:hotelID ?hotelID ;
           ott:hotelLat ?lat ;
           ott:hotelLon ?lon ;
           ott:hotelReviewScore ?score ;
           ott:hasCityParent ?city .
    ?city ott:cityID ?cityID ;
          ott:cityName ?cityName ;
          ott:hasStateParent ?state .
    ?state ott:stateName ?stateName .
}' ')
city_data = []
for hotel, city, state, score, lat, lon in qres:
    if len(city.value) > 0:
        city_data.append([hotel.value, city.value, state.value,
                          score.value, lat.value, lon.value])
city_df = pd.DataFrame(city_data,
                       columns=['Hotel', 'City', 'State', 'Score', 'lat', 'lon'])

dir_path = os.path.dirname(os.path.realpath(__file__))
token_filename = os.path.join(dir_path, 'mapbox_token')
token = open(token_filename).read() # token from Mapbox

import plotly.express as px
fig = px.scatter_mapbox(
    city_df, lat='lat', lon='lon', hover_name='City',
    hover_data=['State', 'Score'], color_discrete_sequence=['fuchsia'],
    zoom=5, height=1000
)
fig.update_layout(mapbox_access_token=token)
fig.update_layout(margin={'r':0, 't':0, 'l':0, 'b':0})
fig.show()

```

Fig. E.8 Example of ontology visualization: Main tourist destinations in Colombia

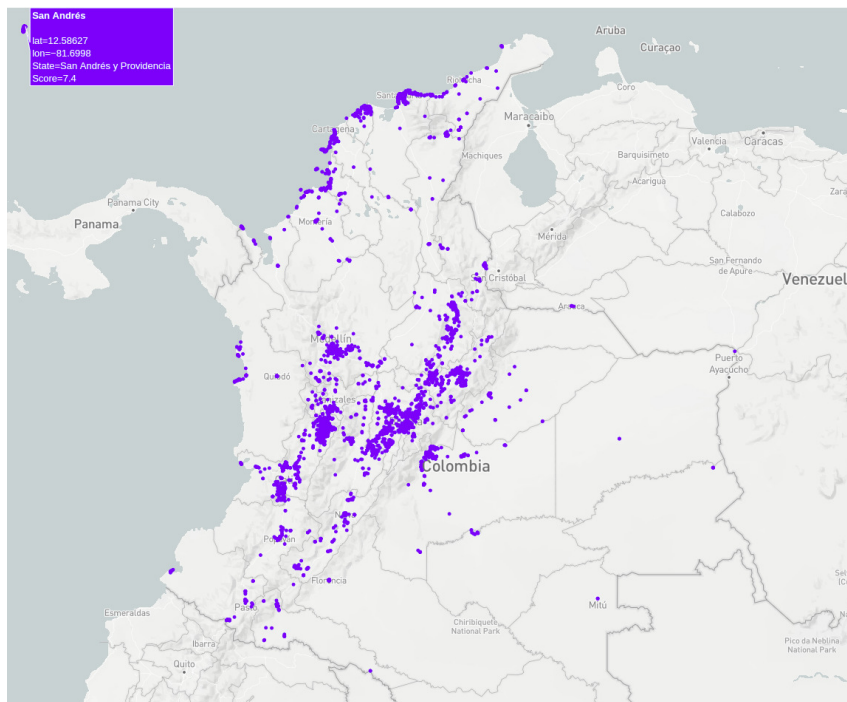


Fig. E.9 Example of the visualization of tourist destinations in Colombia from OntoTouTra.

```

from nltk.sentiment.vader import SentimentIntensityAnalyzer
sia = SentimentIntensityAnalyzer()

reviews['neg'] = reviews['review'].apply(Lambda x:sia.polarity_scores(x)['neg'])
reviews['pos'] = reviews['review'].apply(Lambda x:sia.polarity_scores(x)['pos'])
reviews['neu'] = reviews['review'].apply(Lambda x:sia.polarity_scores(x)['neu'])
reviews['compound'] = reviews['review'].apply(
    Lambda x:sia.polarity_scores(x)['compound'])

from plotly.offline import download_plotlyjs, init_notebook_mode, plot
import plotly.express as px

reviews['YearMonth'] = pd.to_datetime(reviews.date).apply(
    Lambda x: '{year}-{month:02d}'.format(year=x.year, month=x.month))

df = reviews.groupby(reviews.YearMonth)[['pos', 'neg', 'neu']].mean()
df.reset_index(inplace=True)
df = df.rename(columns = {'YearMonth':'date'})
fig = px.line(df, x=df.date, y=[df.pos], hover_data={'date': "|%B %d, %Y"},
    title='Satisfaction KPI')
fig.update_xaxes(dtick="M1",tickformat="%b\n%Y",
    ticklabelmode="period")
plot(fig)

```

Fig. E.10 Application of sentiment analysis techniques to determine the Satisfaction KPI in Colombia

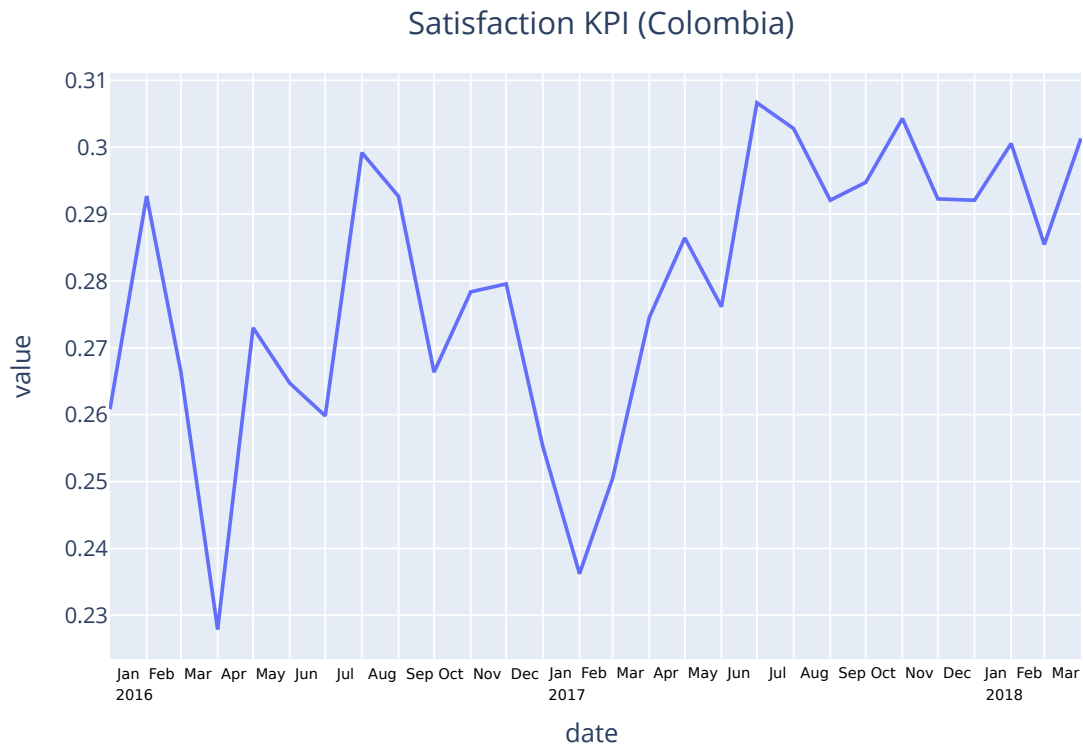


Fig. E.11 Example of satisfaction KPI (Colombia): positive reviews of the destinations. Obtained from OntoTouTra.

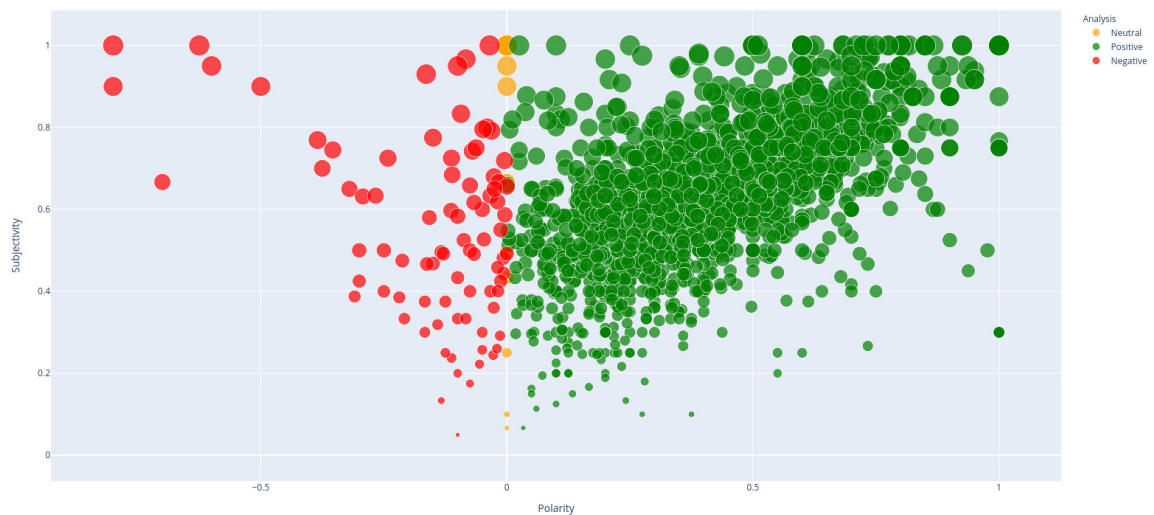


Fig. E.12 Example of the polarity and subjectivity of the reviews about the Colombian destinations. Obtained from OntoTouTra.

```

<div class="bui-grid">
<div class="bui-grid_column-10">
<h3 lang="xu" class=" c-review-block_title c-review_title--ltr ">
Great location , well worth the price
</h3>
</div>
<div class="bui-grid_column-2 bui-u-text-right">
<div class="bui-review-score c-score"> <div class="bui-review-score_badge" aria-label="Scored "> 9.0 </div> </div>
</div>
</div>
</div>
<div class="c-review-block_row">
<div class="c-review">
<div class="c-review_row">
<div class="c-review_inner c-review_inner--ltr">
<p class="c-review_prefix c-review_prefix--color-green"><svg class="bk-icon -iconset-review_great c-review_icon" height="128"
width="128" viewBox="0 0 128 128" role="presentation" aria-hidden="true" focusable="false"><path d="M64 8a56 56 0 1 0 56 56A56 56 0
0 0 64 8zm0 104a48 48 0 1 1 48-48 48 0 0 1-48 48zm44 64a8 8 0 1 1 8-8 8 0 0 1-8 8zm48 8 0 1 1-8-8 8 0 0 1 8 8zm-4.8
21.6a4 4 0 1 .6 3.6A24.3 24.3 0 0 1 64 97c-9.7 0-15.7-4.2-19.7-8a22.7 22.7 0 0 1-4.8-8A4 4 0 0 1 44 76h40a4 4 0 0 1 3.2 1.6z"></
path></svg><span class="bui-u-sr-only">Liked</span> </span><span aria-hidden="true">&nbsp;&nbsp;&nbsp;</span></span><span class="c-
review_body" lang="en-us">Great location , walking distances from several markets , a bakery ,&amp; restaurants. Easy transport
to/from the airport &amp; old/new city . Very friendly and professional staff. Most know some or lots of English which is extremely
helpful . Breakfast is descent , lots of variety to choose from .I would highly recommend .</span>
</p>
</div>
</div>
<div class="c-review_row_lalala">
<p lang="en-us" class="c-review_inner c-review_inner--ltr">
<span class="c-review_prefix"><svg class="bk-icon -iconset-review_poor c-review_icon" height="128" width="128" viewBox="0 0 128
128" role="presentation" aria-hidden="true" focusable="false"><path d="M64 8a56 56 0 1 0 56 56A56 56 0 0 64 8zm0 104a48 48 0 1 1
48-48 48 0 0 1-48 48zm44 64a8 8 0 1 1 8-8 8 0 0 1-8 8zm48 8 0 1 1-8-8 8 0 0 1 8 8zm-5.2 30.2a4 4 0 1 1-5.6
5.6c-10.5-10.4-24-10.4-34.4 0a4 4 0 0 1-5.6-5.6c13.6-13.7 32-13.7 45.6 0z"></path></svg><span class="bui-u-sr-only">Disliked</span>
</span><span aria-hidden="true">&nbsp;&nbsp;&nbsp;</span></span><span class="c-review_body" lang="en-us">My toilet was not cleaned properly .
The room had this weird odor if the AC was turned off</span>
</p>

```

Fig. E.13 Review data stream: unstructured.

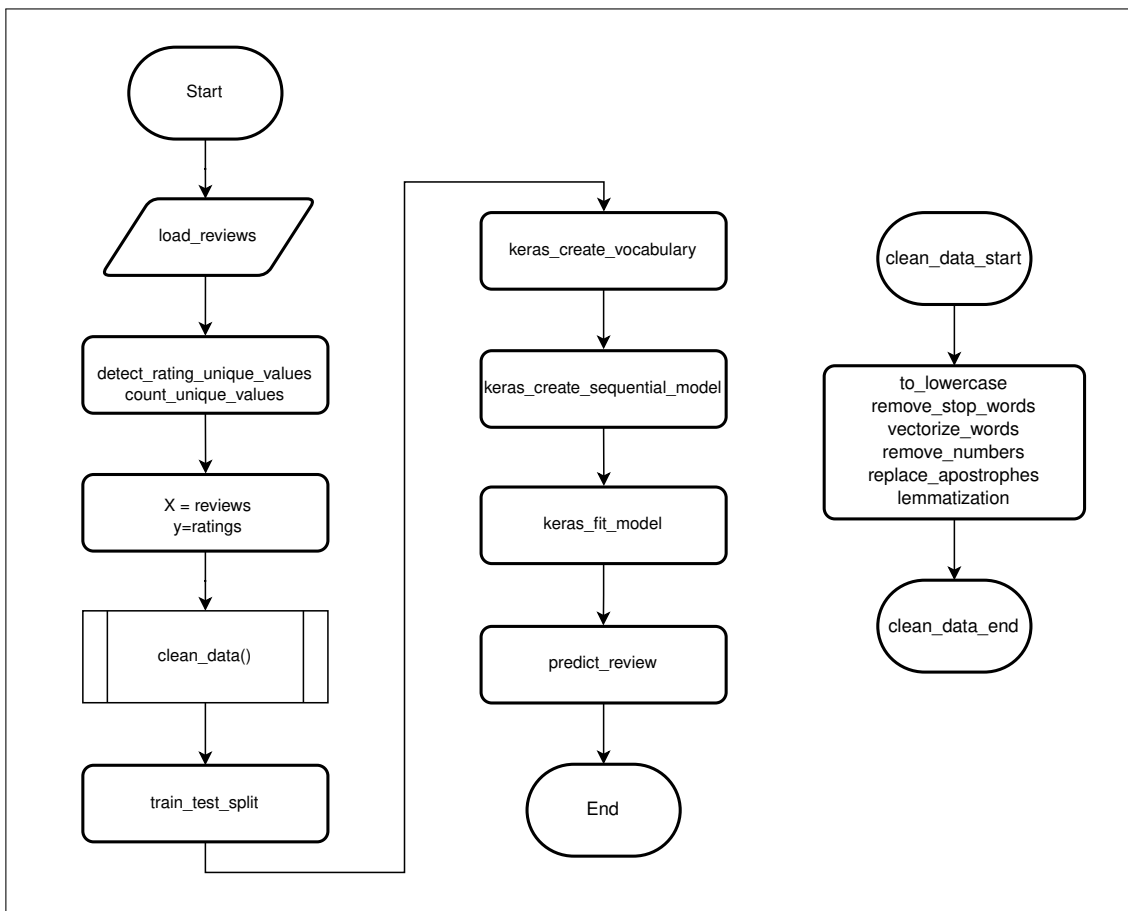


Fig. E.14 Rating predictor algorithm.

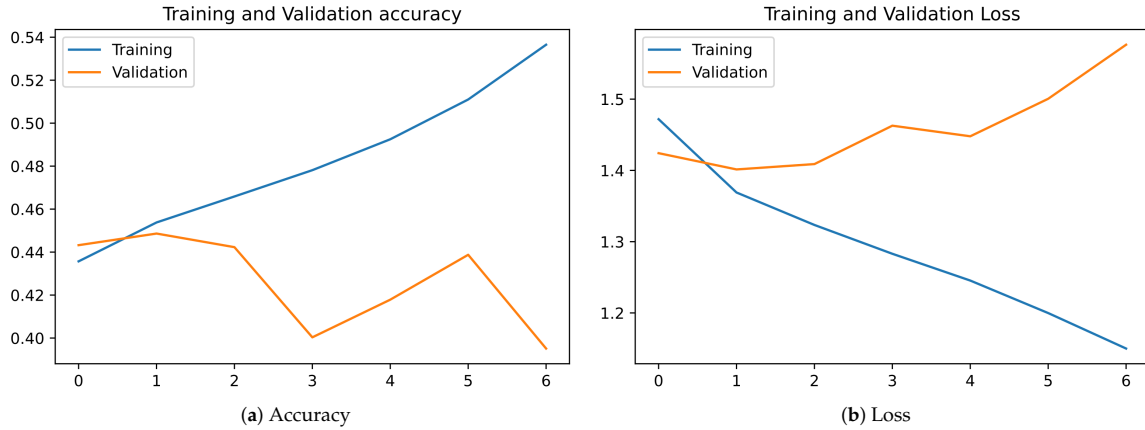


Fig. E.15 Performance of the rating prediction model.

E.1 Data Treatment

This paper presents the methodology of constructing a tourist traceability ontology called OntoTouTra as an educational and research effort. The data to generate the individuals (instances) were obtained from ubiquitous computing sources, especially from social networks, sensors installed in POI, and applications installed on users’ mobile devices. The OntoTouTra ontology, without individuals, and the source code referred to in this paper are available in the repository indicated in Appendix A of this paper.

We can run the source code to obtain the data and feed the ontology with the individuals. Still, before doing this, we strongly recommend that the ToS be reviewed for the data treatment of the owner or owners of these data.

For our case, we reviewed the ToS of Booking.com [179], which was the OTA that we chose to scrape the data to carry out the test cases and the study case. Within these ToS, in the “*Scope & Nature of Our Service*” Section, we find “... *Our Trip Service is made available for personal and non-commercial use only. Therefore, you are not allowed to resell, deep-link, use, copy, monitor (e.g., spider, scrape), display, download, or reproduce any content or information, software, reservations, tickets, products, or services available on our Platform for any commercial or competitive activity or purpose ...*”. On the other hand, in the “*Intellectual Property Rights*” Section, we find: “...*Booking.com exclusively retains ownership of all rights, title and interest in and to (all intellectual property rights of) (the look and feel (including infrastructure) of) the Platform on which the service is made available (including the guest reviews and translated content) and you are not entitled to copy, scrape, (hyper-/deep) link to, publish, promote, market, integrate, utilize, combine or otherwise use the content (including any translations thereof and the guest reviews) or our brand without our express written permission...*”. We can also consult the “robots.txt”

file of the OTA website to verify if it prevents (disallows) crawling or scraping and from the crawl rate to verify if the query is made by a human.

The objective of Krotov and Silva’s research [180, 181] was to identify a set of ethical and legal considerations when collecting data from the web using automated tools. According to them, no legislation directly addresses web scraping. There is a set of theories and laws that guide web scraping, such as “copyright infringement,” “breach of contract” on the side of the web user, the act of computer fraud and abuse (CFAA), and “trespass to chattels. ” In the case of copyrighted material, data that are explicitly owned and copyrighted by the website owner may lead to a case of “copyright infringement.” However, a website does not necessarily own user reviews. Given these conditions, and based on the research reflections, we decided to publish the ontology without the individuals (instances). However, the experimentation environment can be reproducible by feeding this ontology with the data obtained after running the software.