TourMISLOD: a Tourism Linked Data Set

Marta Sabou a,∗, Irem Arsal b and Adrian M.P. Brasoveanu a

a New Media Technology Department, MODUL University Vienna, Austria
E-mail: marta.sabou, adrian.brasoveanu@modul.ac.at

b Tourism and Hospitality Department, MODUL University Vienna, Austria
E-mail: irem.arsal@modul.ac.at

Abstract. The TourMISLOD dataset exposes as linked data a significant portion of the content of TourMIS, a key source of European tourism statistics data. TourMISLOD contains information about the Arrivals, Bednights and Capacity tourism indicators, recorded from 1985 onwards, about over 150 European cities and in connection to 19 major markets. Due to licensing issues, the usage of this dataset is currently limited to the TourMIS consortium, however, a prototype application has already revealed its usefulness for decision support.

Keywords: Tourism indicators, tourism ontology, triplification, decision support

1. Introduction

In this paper we present the TourMISLOD dataset which contains the linked data encoding of European tourism statistics extracted from the TourMIS system. We start by describing the purpose of creation, content and availability of our dataset in Section 2. In Section 3 we provide more details about the dataset creation process and describe current use in Section 4. We conclude with known limitations and planned future work in Section 5.

2. The TourMISLOD Dataset

2.1. Purpose of Creation

The motivation behind creating the TourMISLOD data is threefold. Firstly, data about tourism indicators such as those stored in the TourMIS system are important for tourism decision makers. Secondly, although there are several tourism datasets, none provide the level of details that TourMIS does, nor are they available as linked data. Thirdly, the current technology solutions used by tourism statistic sources are limited to offering data via diverse APIs or simply as data dumps. This severely hampers the development of applications that wish to combine data from different sources. We now detail and exemplify these issues.

The importance of tourism indicators. Tourism indicators play a key role in supporting decision making processes in tourism. The tourism domain is a highly complex and dynamic domain where decision-makers often rely on forecasting models to predict future demand or on decision support systems to analyze and compare the relevant stakeholders (e.g., benchmarking competing regions). Tourism statistics such as the number of tourists that arrive to the destination and the number of bed nights spent at the destination are important for the industry for various decision making related tasks such as (i) understanding the contribution of tourism to the destination’s economy [9] or (ii) promoting and marketing a destination by forecasting tourism demand, setting marketing goals and exploring potential source markets [4]. In addition, tourism planners and public agencies can use tourism statistics to decide on planning tourism related facilities (e.g., hotels and resorts) and infrastructure such as airports, highways, bridges and water treatment facilities [4].

Characteristics of tourism statistics data sources. As a consequence of their importance, many organi-
zations, such as the World Bank\(^3\), the UN\(^4\) or Eurostat\(^5\), provide tourism statistics (see details in [8]). However, none of these datasets provide the level of detail that TourMIS does. While the previously mentioned sources provide annual measurements (except Eurostat), at country level, TourMIS contains both annual and monthly measurements and it focuses on individual cities. Additionally, TourMIS also identifies key markets based on tourists’ origin, a feature not offered by any of the data sources we surveyed, despite the fact that market information is essential for tourism promotion organizations in developing their international advertising campaigns.

Tourism data integration issues. The important activities of decision support often require combining data from various data sources. Indeed, if the decision-maker only makes use of one, isolated, data source his analysis is limited to the data available in that source and ignores other indicators that would allow discovering complex phenomena and designing more accurate forecasting models. However, tourism data sets primarily exist in isolation and they are often difficult to combine and compare automatically. At the data encoding level, while most data sets are published as open data, they use syntactic encoding formats that lead to substantial manual effort when integrating them (e.g., data dumps and in some cases custom APIs). At the data semantics level, they contain data of different geographic granularity, time frequency or they employ different ways of measuring the same indicator, but all these differences are not made explicit in a machine readable format and it remains the task of an analyst to understand them. For example, difficulties caused by this technological status affected BASTIS\(^6\), a system that aims to support tourism decision makers in making better marketing and strategy decisions. BASTIS targets tourism stakeholders involved in heritage tourism in the Baltic Sea region and provides them with information on trends and statistics (both tourism and economic) about this area, thus overcoming the general shortcoming of such information. BASTIS integrates data from TourMIS and Eurostat among others, however, this integration is purely manual and therefore costly and error-prone (based on email communication with the creators of BASTIS).

2.2. Data Source and Coverage

The data source of TourMISLOD is the TourMIS system, an online database that consists of tourism market research data such as bednights, arrivals and capacities in various countries and cities [11]. The major aim of TourMIS is to have comparable data necessary to support tourism managers in their decision-making processes [11]. As such, a supporting consortium, including National Tourism Statistics Austria, European Travel Commission (ETC), European Cities Marketing (ECM) and Austrian National Tourist Office, ensures the continued development and population of the system. TourMIS contains data about three major tourism indicators:

- **Arrivals** - the number of tourists that arrive to various types of accommodations (i.e. hotels, bed and breakfasts, camp sites, etc.) at a destination;
- **Bednights** - the number of nights spent by tourists at various types of accommodations at a destination;
- **Capacity** - the total bed capacity of accommodations at a destination.

This data is provided by several organizations. The National Tourism Statistics Austria collects data from the Austrian accommodation suppliers regarding key tourism indicators. ECM and ETC support the collection of measurements for the three tourism indicators by encouraging their members, city tourism organizations (CTOs) of over 100 European cities and national tourism organizations (NTOs) of 33 nations respectively, to enter their data into TourMIS. The supporting consortium updates the TourMIS data frequently, with new data being added almost daily\(^7\). Data about the three indicators is available from 1985 onwards, in relation with 154 European destinations (i.e., cities) and for 19 different markets, where markets denote the origin of the tourists. The indicators are measured both monthly and annually. Besides storing raw data, TourMIS also includes a method-base that computes a range of statistics such as market shares and market volumes of selected cities.

TourMIS provides a REST API which returns an XML file containing a set of measurements where each measurement is about one of the three tourism indi-

---

\(^3\)data.worldbank.org/
\(^5\)epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database
\(^6\)www.bastis-tourism.info
\(^7\)The latest additions are displayed at www.tourmis.info.
It refers to one destination (e.g., TLL, which is a code for Tallinn), it is about one market, it has an associated year, as well as month if it is a monthly reading, and a value. The following example shows the XML encoding of an arrivals measurement for Tallinn, where 28033 Russian tourists arrived in January 2012.

```xml
<data>
  <destination>TLL</destination>
  <market>RU</market>
  <year>2012</year>
  <month>1</month>
  <value>28033</value>
</data>
```

2.3. Licensing and Availability

Licensing issues are a major issue for our dataset. TourMIS is a system which, although developed at a university, is financed by multiple tourism organizations and is updated by a range of different contributors (Section 2.2). While form-based data extraction from the database is granted for free to anyone upon registration with TourMIS, opening up the entire data set for querying by third parties is a major step that raises intricate licensing issues given the heterogeneous origin of the data. Therefore, for now the linked data we produced remains closed and for use only within the TourMIS consortium, but discussions with the other stakeholders are ongoing about the possibility of opening (at least parts of) this data for public querying. In anticipation of this data being openly accessible, its name already contains “LOD”.

Due to these licensing issues, we are unable to provide this dataset publicly at the moment. We have, however, published a sample of the dataset for inspection by the reviewers of this submission. The sample contains (i) all (1586) Arrivals measurements from 1985 to 2012, measured annually, for all destinations and for the total market (ZZ); (ii) all (9989) Bednights measurements, for all destinations and all markets, measured monthly during 2005; and (iii) all (107) Capacity measurements, for all destinations, for year 2007. While only a portion of the entire dataset, this sample will allow for checking technical correctness as well as it will give an insight into the key characteristics of the dataset (the three measurements, the availability of data over 28 years, for 158 destinations and 19 markets). The dataset is stored in an OpenRDF Sesame repository and can be accessed at tourmislod.modul.ac.at.

3. Linked Data Creation Process

The process of creating the linked data set included three major stages: creating an ontology to represent the concepts covered by TourMIS (Section 3.1), creating RDF data triples from the TourMIS database (Section 3.2) and, finally, establishing links to other data sources (Section 3.3).

3.1. Ontology Creation

We started our linked data creation process with a step involving the identification of existing tourism ontologies that could be used to represent the TourMIS data. Although several tourism ontologies exist their focus is on supporting tourist-centric applications (e.g., recommendation and question answering systems to be used by tourists) and their vocabulary is restricted by those applications’ scope [1]. For example, QALL-ME provides a model to describe tourism destinations, sites, events as well as transportation [7]. The Harmonise ontology focuses on tourism events and accommodation types [3], while the Hi-Touch ontology models tourism destinations and their associated documentation [5]. We were unable to identify any ontology providing concepts about the tourism indicators of interest to us and therefore created our ontology as we describe next.

Figure 1 depicts the ontology that models the various measurements and their characteristics. Central to the ontology is the Measurement concept and its three subconcepts that correspond to the tourism indicators covered by TourMIS. This concept also serves as the domain for a set of object and data properties used to define the key elements of each measurement. The three data properties (depicted with dashed arrows in Figure 1) allow specifying the value of a measurement (hasValue), as well as the year and the month when it was measured (hasYear and hasMonth respectively). The object properties (depicted with dotted arrows in Figure 1) allow specifying the Destination and the Market for which the measurement has been made as well as any difference in measurement with respect to the definition of the respective measurement.
3.2. Triplification

We triplified a subset of TourMIS containing raw statistical data about the Arrivals, Bednights and Capacity tourism indicators. The data spans 28 years, 154 destinations and 19 markets (as well as three generic markets that cover the domestic, the foreign and the total market). For transforming the TourMIS content into RDF data based on the ontology, we extracted this data using the server’s REST API. This data was then transformed into an ontology model using the Jena library and saved into RDF files. We have triplified a total of 201,762 measurements with the final data set accounting to just over 1 million triples.

3.3. Creating Links to Other Datasets

To lift the triplified dataset to a 5-star linked data quality level, we have established links between DBpedia resources and the corresponding destinations in TourMIS (154 European cities) as well as the 19 countries that constitute the key markets covered by the system. Links for both cities and countries were identified by querying DBpedia for entities that (i) were of type dbo:PopulatedPlace and (ii) had the same English label as the label of the city/country in TourMIS (for each code used, such as TTL or RU, TourMIS provides a corresponding label). With this query we suc-
cessfully linked all countries to the corresponding DBpedia entity, and all except 20 cities. The major reason for failing to find a link, was, in most of the case, that a city was not of a dbo:PopulatedPlace type. Given the small number of outliers, we preferred to manually add the correct links for these cities instead of defining another query that would successfully identify links for them.

We also established a schema level mapping by specifying that the Country concept is the same as http://dbpedia.org/ontology/Country.

4. Current Usage

The TourMISLOD dataset has been created in March 2012 and has not been made publicly available due to the licensing issues described in Section 2.3. As a result, current usage is limited to efforts internal to the MODUL university, where we explore the potential of this dataset in supporting decision making processes. Decision support is a task of major interest to members of the TourMIS consortium as well as to the typical users of the system. Indeed, TourMIS caters for four main user groups: firstly, representatives of national, provincial, regional and city tourism organizations, which are involved in long-term, strategic planning of the tourism development of a region; secondly, tourism suppliers such as suppliers of accommodation, food, travel, culture, sport as well as travel agencies and tour operators, which are mostly interested in local forecasts; thirdly, educational institutions active in tourism research and fourthly, consultants and public authorities involved in regional planning and decision making. Typical tourism decision making processes as well as the benefit of using linked data to support them are detailed in [8].

We are currently extending the prototype system built on TourMISLOD and described in [8] with functionalities that allow drill down analysis of the dataset, a common feature of decision support systems [2]. Concretely, our initial prototype allowed a graphical comparison between TourMIS and World Bank arrivals data for a given European country. The new application also displays additional graphs showing the destinations from which data was collected for a given time-point as well as the relevant markets. This allows a comparison between the tourist numbers visiting each destination in a country, as well as an insight into the markets that visit each destination. The prototype is available at tourmislod.modul.ac.at.

5. Conclusions

To conclude, we discuss shortcomings of the data set and envisioned future work.

5.1. Shortcomings of TourMISLOD

A key shortcoming of TourMISLOD is its currently restrictive license due to the heterogeneous origin of the source data (see Section 2.3). We hope that this restriction will be lifted soon, pending the decision of the TourMIS consortium. Meanwhile, requests from interested developers will be treated on a case by case basis, where providing part of the dataset might be possible.

Another shortcoming is that the current dataset creation process requires the linked data set to be regenerated in order to include updates. This solution does not suit the dynamic nature of the TourMIS data as new updates are not accessible as Linked Data as soon as they are added to TourMIS but only when the data set is regenerated. As a possible solution, we have investigated the use of existing database to RDF translators such as D2R or other tools providing RDF/SPARQL access to relational databases, however, we did not use them at this stage because the TourMIS database itself was undergoing a re-design to ensure its scalability. Adopting such an automated solution once TourMIS has been redesigned is future work.

5.2. Future Work

Future work will focus, firstly, on exposing not just raw data, but also data points derived by the model base such as market-size, market-share or forecasted values. Here careful considerations must be given to correctly conveying the meaning of the statistical formula used to derive these data points, as described in [10]. Secondly, we will also expose data about tourism sights available in TourMIS. This will provide an additional dimension to the data, and will raise more complex linking problems. It will also require an extension of the current ontology to cover new types of tourism concepts, potentially by reusing existing tourism domain models (see Section 3.1).

9About two months before the time of writing.

10Such as those listed at d2rq.org/resources
In terms of applications on top of this dataset, we will extend the current prototype in line with requirements from the TourMIS consortium and to include additional data sources, more indicators and diverse visual metaphors (e.g., maps). We will also explore collaboration with projects that currently use a manual approach to re-use TourMIS data, such as BASTIS.

Acknowledgements

The work presented in this paper was developed within DIVINE (www.weblyzard.com/divine), a research project funded by FIT-IT Semantic Systems of the Austrian Research Promotion Agency (www.ffg.at) and the Federal Ministry for Transport, Innovation and Technology (www.bmvit.gv.at).

References


