Undefined 1 (2009) 1–5
IOS Press

TourMISLOD: a Tourism Linked Data Set

Editor(s): Pascal Hitzler, Wright State University, USA; Krzysztof Janowicz, Pennsylvania State University, USA
Solicited review(s): Amit Joshi, Wright State University, USA; Michael Lutz, European Commission -Joined Research Centre, Italy; Jesse Weaver, Rensselaer Polytechnic Institute, USA

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. Nevertheless, a prototype application has already revealed the dataset's usefulness for decision support.

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

1. Introduction

We present the TourMISLOD[1] dataset which contains the linked data encoding of European tourism statistics extracted from the TourMIS[2] system. We start by describing the content, the purpose of creation and availability of our dataset in Section 2. In Section 3 we provide more details about the dataset creation process and describe current usage in Section 4. We discuss limitations and future work in Section 5.

2. The TourMISLOD Dataset

2.1. 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 European countries and cities [12]. Such data collections are referred to as *tourism indicators* or *tourism statistics*. The major aim of TourMIS is to have comparable data necessary to support tourism managers in their decision-making [12]. 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 organiza-

*Corresponding author. E-mail: marta.sabou@modul.ac.at.
[1] Available at `http://tourmislod.modul.ac.at/`
[2] `www.tourmis.info/index_e.html`

tions (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[3]. Data about the three indicators is available from 1985 onwards, in relation with 154 European destinations[4] (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 indicators, it refers to one destination (e.g., LJU, which is a code for Ljubljana), it is about one market, it has an associated year, as well as month if it is a monthly reading, and a value. The example in Listing 1 shows the XML encoding of a bednights measurement for Ljubljana, where 837 bednights were spent by Spanish tourists in March 2005.

Listing 1: XML output from the TourMIS API

```
<data>
    <destination>LJU</destination>
    <market>ES</market>
    <year>2005</year>
    <month>3</month>
    <value>837</value>
</data>
```

2.2. 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 [10] 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 organizations, such as the World Bank[5], the UN[6] or Eurostat[7], 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 relies on one, isolated, data source his analysis ignores other, external 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 because

[3] The latest additions are displayed at www.tourmis.info.

[4] http://tourmislod.modul.ac.at/coveragemap

[5] http://data.worldbank.org/

[6] http://data.un.org/

[7] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

of the following two issues. Firstly, 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). Secondly, 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[8], 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 but this integration is purely manual and therefore costly and error-prone (based on email communication with the creators of BASTIS).

2.3. Licensing and Availability

Licensing is 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 2.1). 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. For now, it has been agreed that partial access (e. g., to data collected before 2010) or full access to the data might be granted upon request. 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 interested readers. 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[9] and can be accessed at `http://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), dataset publication (Section 3.3) and, finally, establishing links to other datasources (Section 3.4).

3.1. Ontology Creation

We started with the identification of 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 documentations [5]. We were unable to identify any ontology defining the tourism indicators of interest to us and therefore created our ontology as described 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

[8] `http://www.bastis-tourism.info`

[9] `http://www.openrdf.org`

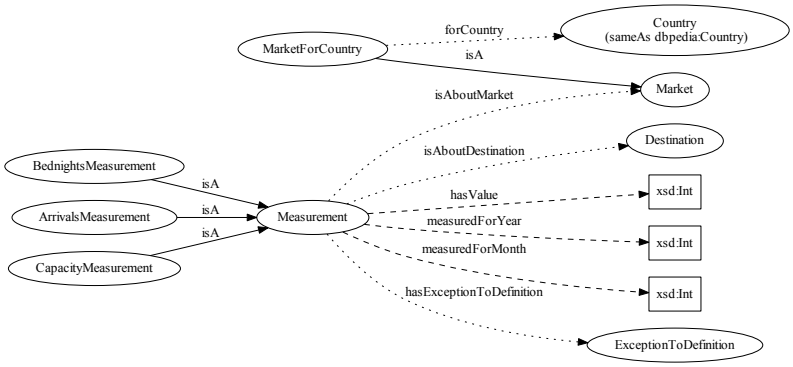

Fig. 1. Ontology schema for describing key TourMIS concepts (xsd:Int stands for xsd:integer).

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 doted 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 that measurement type (`hasExceptionToDefinition`). The `ExceptionToDefinition` concept records, using a textual comment, cases when the measurement differs from the main definition, for example, when arrivals are measured "in city area only" or "in greater city area". Currently, we model these exceptions as TourMIS does, i.e., as textual comments, however, we envision a more formal, axiom-based modeling as future work. When specifying a market for a measurement, one can either refer to three generic markets (total, domestic and foreign markets codified as ZI, ZA and ZZ respectively) or to markets specific to a given country. A market for a specific country is an instance of `MarketForCountry` and specifies the country of interest through the `forCountry` property.

During ontology design we relied on the "N-ary relation with no distinguished participant" design pattern recommended by [6] for modeling relations between many individuals, where none of the individuals can be considered a primary one. By using this design pattern we closely follow the structure of the TourMIS XML output, thus making the triplification process, as well as debugging activities more straightforward.

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 (and three generic markets). 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. In Listing 2 we provide the RDF representation of the data from Listing 1. To assess internal connectivity, we have computed the average ratio of incoming vs. outgoing links for each node in the RDF data. We obtained a value of 2.9.

3.3. URI Design and Data Set Publication

We use different namespaces for ontology concepts and for instances, namely `http://tourmislod.modul.ac.at/tourmis/ontology/` and `http://tourmislod.modul.ac.at/tourmis/resource/` respectively. In Listing 2 and in the Sesame repository, we refer to these namespaces as *to* and *tr* respectively. Each measurement is assigned a generated local name which contains its type, the TourMIS code for the destination, the year and month of its measurement as well as the TourMIS code for the respective market, if any, as exemplified in Listing 2. We make use of slash URIs which have the expected 303 redirect behavior [9].

We have used the Pubby[10] tool to add a Linked Data interface of dereferenceble URIs on top of Sesame's

[10] `http://www4.wiwiss.fu-berlin.de/pubby/`

Listing 2: The RDF representation of the data from Listing 1.

```
<rdf:Description rdf:about="&tr;aBednightsMeasurement_LJU_2005_3_ES">
  <to:isAboutMarket rdf:resource="&tr;marketForSpain"/>
  <to:measuredForMonth>3</to:measuredForMonth>
  <to:measuredForYear>2005</to:measuredForYear>
  <to:hasValue>837</to:hasValue>
  <to:isAboutDestination rdf:resource="&tr;Ljubljana"/>
  <rdf:type rdf:resource="&to;BednightsMeasurement"/>
</rdf:Description>
```

SPARQL endpoint. A good starting point for exploring the dataset is `http://tourmislod.modul.ac.at/tourmis/ontology/Measurement`. Since this page corresponds to the core concept of our ontology, it provides access to all schema elements as well as to all measurement instances.

3.4. Creating Links to Other Datasets

To lift the triplified dataset to a 5-star linked data 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 LJU or ES, TourMIS provides a corresponding label). With this query we successfully 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 manually added the correct links for these cities. We established a schema level mapping by specifying that the `Country` concept is the same as `http://dbpedia.org/ontology/Country`.

We also extended the dataset with links to the GeoNames[11] geographic dataset for the cities. Most of these links were identified from DBpedia for those cities whose DBpedia entry contained a GeoNames link. For 22 cities that did not contain any GeoNames links, we looked up the corresponding URIs manually. To sum up, external connectivity amounts to 328 `owl:sameAs` links to two external datasets: DBpedia (174) and GeoNames (154).

4. Current Usage

The TourMISLOD dataset has been created in March 2012[12] 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 domestic tourism demand; 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]. Our initial prototype allowed a graphical comparison between TourMIS and World Bank arrivals data

[11] `http://www.geonames.org/`

[12] About two months before the time of writing.

for a given European country. The new application also displays additional graphs showing the destinations from which data was collected for a given timepoint. This allows a comparison between the tourist numbers visiting each destination in a country. The prototype is available at `http://tourmislod.modul.ac.at`.

5. Conclusions

A shortcoming of TourMISLOD is its currently restrictive license due to the heterogeneous origin of the source data (see Section 2.3). While the TourMIS consortium decides on the appropriate license, requests from interested developers will be treated on a case by case basis, where providing part of the (or the entire) dataset might be possible. Beyond our project, providing licensing agreements for datasets where the data belongs to multiple stakeholders, is an interesting future research issue relevant for the entire field.

Another issue 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 only accessible as Linked Data 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[13]: 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.

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 [11]. 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).

[13] Such as those listed at `http://d2rq.org/resources`

In terms of applications, we will extend the current tool 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 adopt a manual re-use of TourMIS data, such as BASTIS.

Acknowledgements

We thank Prof. Karl Wöber for his support with TourMIS and for communicating with the TourMIS consortium on our behalf. 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

[1] R. Barta, C. Feilmayr, B. Pröll, C. Grün, and H. Werthner. Covering the Semantic Space of Tourism: An Approach based on Modularized Ontologies. In *Proc. of the 1st WS. on Context, Information and Ontologies*, CIAO '09, pages 1–8. ACM, 2009.

[2] F. Burstein and W. C. Holsapple, editors. *Handbook on Decision Support Systems*. Springer, 2008.

[3] O. Fodor and H. Werthner. Harmonise: A Step Toward an Interoperable E-Tourism Marketplace. *Int. J. Electron. Commerce*, 9(2):11–39, January 2005.

[4] D. C. Frechtling, editor. *Forecasting Tourism Demand: Methods and Strategies*. Butterworth Heinemann, 2001.

[5] Mondeca. Semantic Web Methodologies and Tools for Intra-European Sustainable Tourism. White Paper, 2004.

[6] N. Noy and A. Rector. Defining N-ary Relations on the Semantic Web. W3C Working Group Note, 2006.

[7] S. Ou, V. Pekar, C. Orasan, C. Spurk, and M. Negri. Development and Alignment of a Domain-Specific Ontology for Question Answering. In *Proc. of the Sixth International Language Resources and Evaluation Conf. (LREC)*, 2008.

[8] M. Sabou, A. Brasoveanu, and I. Arsal. Supporting Tourism Decision Making with Linked Data. In *Proc. of the of the 8th Int. Conf. on Semantic Systems (I-SEMANTICS)*, 2012.

[9] L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web. W3C Interest Group Note, http://www.w3.org/TR/cooluris/, 2008.

[10] W. B. Stronge. Statistical Measurements in Tourism. In *VNR's Encycl. of Hospitality and Tourism*, pages 735 – 745. 1993.

[11] D. Vrandecic, C. Lange, M. Hausenblas, J. Bao, and L. Ding. Semantics of Governmental Statistics Data. In *Proc. of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.

[12] K. Wöber. Information supply in tourism management by marketing decision support systems. *Tourism Management*, 24(3):241 – 255, 2003.