Supporting Tourism Decision Making with Linked Data

Marta Sabou
New Media Technology Department
MODUL University Vienna
marta.sabou@modul.ac.at

Adrian M.P. Brașoveanu
New Media Technology Department
MODUL University Vienna
adrian.brasoveanu@modul.ac.at

Irem Arsal
Tourism and Hospitality Department
MODUL University Vienna
irem.arsal@modul.ac.at

ABSTRACT
Decision makers in the tourism domain routinely need to combine and compare statistical indicators about tourism and other related areas (e.g., economic). While many organizations offer relevant data sets, their automatic access and reuse is hampered (i) by them being offered as data dumps in non-semantic encodings; (ii) by them assuming some implicit knowledge that is necessary to build applications (e.g., that a city is situated in a certain country) and (iii) by the use of incompatible ways to measure the same indicator without formally specifying the assumptions behind the measurement technique. We explore the use of linked data technologies to solve these issues by triplifying the content of TourMIS, a broadly used data source of European tourism statistics and by building a prototype system using this data.

Keywords
Tourism, Tourism Indicators, Tourism statistics, Triplification, Data Visualization

1. INTRODUCTION
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., competing regions). Tourism statistics such as the number of tourists that arrive to and sleep at a 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 [3] or (ii) promoting and marketing a destination by forecasting tourism demand, setting marketing goals and exploring potential source markets [2]. In addition, tourism planners and public agencies can use tourism statistics to decide on planning tourism related facilities and infrastructure such as airports, highways, bridges and water treatment facilities [2]. These important activities 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.

As a consequence of their importance, many organizations, such as the World Bank, the UN or Eurostat, provide tourism statistics (see Section 2.1). However, these tourism related data sets primarily exist in isolation and they are often difficult to compare, as they contain data of different geographic granularity, time frequency or they employ different ways of measuring the same indicator. While most data sets are published as open data, they use syntactic encoding formats that lead to substantial manual effort when integrating them. For example, difficulties caused by this technological status affected BASTIS[1], 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 (a key source of European tourism statistics detailed in Section 2.2.) and Eurostat among others, however, this integration is purely manual and therefore very costly and error-prone (based on email communication with the creators of BASTIS).

To overcome such situations, we propose exposing the content of TourMIS as Linked Data and explore its combination with data from other sources to support typical decision-making scenarios such as those described in Section 2. While the tourism domain is often seen as a key application domain for linked (open) data technologies, typical tourism applications that make use of linked data have mostly focused on the needs of tourists (e.g., DBpedia Mobile). In contrast, the needs of tourism decision makers have been ignored so far, although they routinely need to combine and make sense of large, distributed and heterogeneous data sets.

We continue by describing the details of our use case in Section 2, then we provide technical details about the data triplification and interlinking activities in Sections 3 and 4 respectively. We describe a prototype application that makes use of the created linked data in Section 5, we discuss the benefits provided by linked data to this use case in Section 6 and conclude in Section 7.

2. USE CASE
Our use case focuses on the needs of decision makers in tourism, and in particular on enabling the following two scenarios that they frequently encounter. Firstly, given the statistic (and therefore often unpredictable) nature of the tourism industry, when trying to understand the evolution of tourism indicators (over time, across competing regions, etc), decision makers need to combine tourism related data from different, complementary data sources which might provide different granularity and coverage for that indicator. Secondly, when forecasting future demand, decision-makers must broaden their investigations to include other indicators besides those in the tourism area and try to discover any

I-SEMANTICS 2012, 8th Int. Conf. on Semantic Systems, Sept. 5-7, 2012, Graz, Austria

[1] http://www.bastis-tourism.info

significant correlations thereof. For example, economic indicators such as inflation rate[2] or unemployment rate[3] can have a negative effect on tourism. Or, when exploring the environmental effects of tourism, a decision maker must investigate any correlations that might exist between tourism statistics and sustainability indicators such as the percentage of forest area in a country[4]. We continue with an overview of key data sources for tourism decision makers and then provide details about the TourMIS system.

2.1 Tourism Data Sets

The UN provides a multitude of datasets from its offices[5], including also various tourism indicators from UN's World Tourism Organization (UNWTO), such as, among others, arrivals, departures and tourist expenditures, measured per country and year. Much of this data is open, however, only provided in the proprietary Excel format. Additional UNWTO data is only available as pdf downloads[6].

Eurostat provides a wealth of European level statistics[7], including various tourism indicators (capacity, arrivals, bednights, expenditures etc) for all European countries. Measurements are provided on a monthly basis. Data can be downloaded in a variety of formats including Excel, CSV, HTML, SPSS and PDF.

The World Bank provides open access to over 8000 indicators[8], including also a variety of tourism indicators measured annually and at a country level. The available datasets can be downloaded in XML and Excel formats, as well as accessed through an API. There are some third party efforts to provide this data as linked data such as those coordinated by ESDS[9].

We conclude that there are various sources that offer a multitude of indicators in the area of tourism and beyond. A general trend is offering this data as open data, primarily through downloading it in popular (non-semantic) encodings. We foresee however a next stage when more thought will be given to adding metadata to these datasets as well as interlinking them in the spirit of the linked data movement and as pointed out by a recent blogpost at the World Bank [1]. Through the work reported here, we take a step in this direction with the TourMIS system.

2.2 The TourMIS System

The TourMIS system[10] is an online database that consists of tourism market research data such as bednights, arrivals and capacities in various countries and cities [5]. The major aim of TourMIS is to have comparable data to help tourism managers in their decision-making processes [5]. 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 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 and institutions.

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 bednights spent at various types of accommodations;
- Capacity – the total bed capacity at 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[11]. Data about the three indicators is available from 1985 onwards, in relation with 154 European destinations (i.e., cities) and for 19 different markets. The indicators are measured both monthly and annually. As such, this dataset is more detailed than similar tourism related datasets (Section 2.1). While other 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, although market information is essential for tourism promotion organizations in developing their international advertising campaigns. Besides storing raw data, TourMIS includes a method-base that computes a range of statistics such as market shares and market volumes of selected cities.

3. 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 and are codified as ZI, ZA and ZZ respectively).

TourMIS's REST API 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., TLL, which is a code for Tallinn), it has an associated year, as well as month if it is a monthly reading, and a value. The Arrivals and Bednights measurements also specify a market denoting the country from where the tourists come from (e.g., RU, for Russia). The following example shows the XML encoding of an arrivals

[2] http://data.worldbank.org/indicator/FP.CPI.TOTL.ZG

[3] http://data.worldbank.org/indicator/SL.UEM.TOTL.ZS

[4] http://data.worldbank.org/indicator/AG.LND.FRST.ZS

[5] http://data.un.org/Explorer.aspx

[6] http://www.unwto.org/facts/menu.html

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

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

[9] http://www.esds.ac.uk/international/access/LDaccess.asp

[10] http://www.tourmis.info

[11] See http://www.tourmis.info for the latest TourMIS updates.

measurement for Tallinn, where 28033 Russian tourists arrived to Tallinn in January 2012. We have triplified a total of 201 762 measurements.

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

For the triplification process, we designed an ontology to describe the various measurements and their characteristics (Figure 1). The ontology models the various types of measurements and concepts needed to define their properties. `ExceptionToDefinition` is used to record, using a textual comment, cases when the measurement differs from the main definition, for example, when arrivals are measured "in city area only (AG)" or "in greater city area (AGS)". 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.

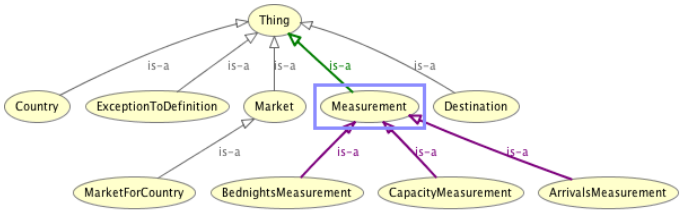

Figure 1: Concept Hierarchy for the TourMIS Ontology

For transforming the TourMIS content into RDF data based on the ontology, we extracted the data described above using the server's REST API. This data was then transformed into an ontology model using the Jena[12] library and saved into an RDF files.

We have also considered 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. We will consider adopting such an automated solution once TourMIS has been updated, as it will better suit the dynamic nature of the TourMIS data and it will allow new updates to be accessible as Linked Data as soon as they are added to TourMIS (with the current solution, the linked data set has to be regenerated to include updates).

The final data set accounts to just over 1 million triples. Due to licensing issues described in Section 7, we are unable to provide this dataset publicly. 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[14] and can be accessed from the resource page dedicated to this paper at http://tourmislod.modul.ac.at/tourmis.html.

4. INTERLINKING

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 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) and, additionally, they were of type `dbo:PopulatedPlace`. 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 did not have a `dbo:PopulatedPlace` property. 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.

6. APPLICATION

While our efforts so far have been focused on triplifying and publishing (internally) the TourMIS dataset, we have also started to develop applications that take advantage of the benefits of the linked dataset. A first prototype[15], depicted in Figure 2, allows a visual comparison between data from TourMIS and other sources (currently from the World Bank). The interface allows selecting a country of focus and a relevant indicator (currently we only support the Arrivals indicator, but we are developing support for the others as well). Since TourMIS offers city level data, we add up all the data from the cities belonging to the selected country and visualize it against country-level statistics from the World Bank. However, TourMIS does not contain any information about the country to which a destination belongs. To solve this issue, we query DBPedia for this information and therefore automatically determine the destinations that are relevant for a given country.

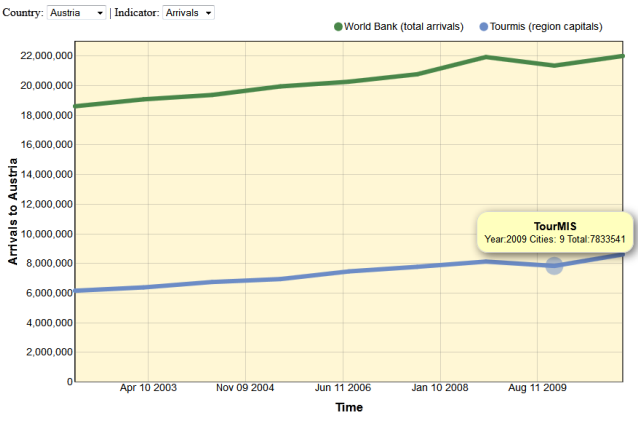

Figure 2:Prototype Interface

The prototype supports decision makers in the first scenario described in Section 2, in which they need to compare similar

[12] http://incubator.apache.org/jena

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

[14] http://www.openrdf.org

[15] Also available at http://tourmislod.modul.ac.at/tourmis.html.

tourism datasets which have a different granularity (in our case, a geographic granularity). The visualization supports a variety of analysis tasks. Firstly, the graph illustrates the share of tourism arrivals in the 9 main Austrian cities with respect to arrivals in the entire country. According to this graph, cities attract only one third of Austrian tourism, but almost the entire growth in the industry is due to intensified city tourism. This could lead to decisions related to intensifying the marketing for city or general tourism. Secondly, the similar trends of the two lines (e.g., growth of arrivals over years, quick recovery after the 2011 crisis period) prove that the two data sets correlate well and that they are both correct. Finally, by taking advantage of the higher granularity of TourMIS data, this prototype could allow "diving into details", by showing how the total arrivals provided by TourMIS are distributed over the 9 Austrian cities and the 19 markets. This ability to zoom into details is a key feature of all decision support systems, and will be implemented as future work.

7. BENEFITS OF USING LINKED DATA

We hereby discuss the benefits of linked data technologies for the tourism domain and explore possibilities of evaluating these benefits. A first benefit refers to syntactic interoperability, as linked data technologies could enable a common technical infrastructure for sharing tourism data that would go beyond the current practice of providing data dumps and APIs, and, consequently facilitate the automatic and dynamic consumption of these data sets in novel (decision supporting) applications. Similarly to the case of scientific data, LD could provide a technical solution for exposing data in a uniform format, allowing easy access to it and facilitating cross-source data integration [6].

Another benefit comes from the links that are established between data sets. While these links facilitate data integration primarily, they also allow augmenting one data source with additional knowledge. For example, by linking TourMIS cities to the corresponding DBpedia resources, we could automatically detect the country where each destination is. As such, within this prototype, we augmented TourMIS with this location information, which it does not provide. Without this information, leveraged thanks to the previously established links, detecting which cities belong to which countries would have been a tedious manual task.

The semantics that linked data sets can carry is an additional benefit. For example, using formal structures to specify the definition of each measurement (and exceptions to it) as opposed to simply adding a text comment, would allow comparing the definition of measurements within and across data sets thus making sure that the right data are compared. Solutions such as using OpenMath [4] will be investigated as future work.

Evaluating the benefits of the proposed linked data based solution with concrete, quantifiable measures is hampered by the following factors. Firstly, we just completed the data triplification stage and only had time to build one prototype application. As with many linked data sets, their benefits are only understood after a period of time needed by the community to build innovative applications with it. Therefore, we plan to measure success as directly proportional to the adoption of this dataset over time by the TourMIS consortium and the public (subject to the licensing issues detailed in Section 7). Secondly, the proposed LD solution aims to support the decision making process, whose benefits are by large intangible. We can however indirectly measure the success of our project by the number and complexity of decision processes (e.g., structured, semi-structure and unstructured decisions) that can be supported.

8. SUMMARY AND OUTLOOK

We have focused on the use case of supporting tourism decision makers in their activities of combining and comparing statistical indicators. After transforming TourMIS data into linked data and building a system that uses it we concluded that linked data technologies (i) can support uniform data access; (ii) allow bringing in additional knowledge from third party sources thanks to the established links; and (iii) can be used to formally specify the measurements used by different sources thus allowing for their automatic combination and comparison (this is future work).

Licensing has been a major issue for our project. TourMIS is co-financed by multiple organizations and is updated by a range of different contributors (Section 2.2). While form-based data extraction is free 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 data's heterogeneous origin. Therefore, for now the linked data remains closed and for use only within the TourMIS consortium, but discussions with the other stakeholders of the system are ongoing about the possibility to open (at least parts of) this data for public querying.

We have many future plans. Firstly, we will continue by exposing not just raw data, but also data points derived by the model base such as market-size or market-share. Here careful considerations must be given to conveying the meaning of the statistical formula used to derive these data points, as described in [4]. Secondly, we will also expose data about sights/attractions from TourMIS. This will provide an additional dimension to the data, and will raise more complex linking issues. In terms of applications, 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 and integrate TourMIS data, such as BASTIS.

9. ACKNOWLEDGMENTS

This work was developed within DIVINE, a project funded by *FIT-IT Semantic Systems* of the *Austrian Research Promotion Agency* (FFG) and the *Federal Ministry for Transport, Innovation and Technology (BMVIT)*. A. Braşoveanu was partially supported by the strategic grant POSDRU/88/1.5/S/60370 (2009) on "Doctoral Scholarships" of the Ministry of Labor, Family and Social Protection, Romania, co-financed by the European Social Fund – Investing in People.

10. REFERENCES

[1] Badiee, S. 2012. Open Data at the World Bank: 2 years today. Blogpost. http://blogs.worldbank.org/opendata/open-data-at-the-world-bank-2-years-old-today

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

[3] Stronge, W. B. 1993. Statistical Measurements in Tourism. In VNR's Encycl. of Hospitality and Tourism, pp. 735-745.

[4] Vrandecic, D., Lange, C., Hausenblas, M., Bao, J. and Ding, L. 2010. Semantics of Governmental Statistics Data. In *Proc. of the WebSci*.

[5] Wöber, K. 2003. Information supply in tourism management by marketing decision support systems. *Tourism Management*. 24:3, pp:241-255.

[6] Zapilko, B., Harth, A. Mathiak, B. 2011. Enriching and analysing statistics with Linked Open Data. In Eurostat Conf. on New Techniques and Technologies for Statistics (NTTS)