Undefined 1 (2009) 1–5
IOS Press

Visualizing Statistical Linked Knowledge Sources for Decision Support

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Adrian M.P. Braşoveanu [a,*], Marta Sabou [b], Alexander Hubmann-Haidvogel [a], Daniel Fischl [a] and Arno Scharl [a]

[a] *New Media Technology Department, MODUL University Vienna, Am Kahlenberg 1, 1190 Vienna, Austria*
E-mail: {adrian.brasoveanu, alexander.hubbman, daniel.fischl, arno.scharl}@modul.ac.at
[b] *Christian Doppler Laboratory for Software Engineering Integration for Flexible Automation Systems, Vienna University of Technology, Favoritenstrasse 9-11/E188, A-1040, Vienna, Austria*
E-mail: {marta.sabou}@ifs.tuwien.ac.at

Abstract. Most Decision Support Systems (DSS) are tailored towards specific domains and use relevant information for certain types of decisions. In today's interconnected world, enriching DSS with external data about events such as financial crises and climate change can improve the decision-making process. One method to build DSS tools that leverage such cross-domain information is to look at the summary of these events as expressed through statistical data. Following the RDF Data Cube (QB) standard there was an increase in the publication of such data and related visualizations, but less effort was dedicated to integrating visualizations into analytical platforms to answer complex questions. After reviewing the relevant work in the field of Linked Data Visualization, this paper describes: (i) a methodology to integrate cross-domain statistical data sources by applying selected QB principles (observations and slicing, for example) to a visual dashboard; (ii) a set of visualization scenarios for cross-domain datasets from multiple sources including Eurostat and the World Bank; and (iii) a dashboard prototype developed following these principles and scenarios.

Keywords: Linked Data, Information Visualization, Decision Support Systems, RDF Data Cube, Tourism

1. Introduction

Most Decision Support Systems (DSS) in use today are tailored towards specific domains and use relevant information for certain types of decisions needed in that domain. In today's interconnected world this is not enough, as external events such as financial crises or climate change (mainly through its consequences: hurricanes, ice meltdown, drought, rising temperatures, and so on) can make such domain specific systems age faster than they should. Building full fledged monitoring systems into any DSS is a good method through which we can address some of the mentioned problems, but unfortunately one that is currently expensive and out-of-reach for most research groups and companies. One method to build DSS tools that leverage such cross-domain information is to look at the summary of these events as expressed through statistical data. Indeed such an operation would help significantly improve current systems while also offering as additional benefit the possibility to answer complex questions that would require cross-domain data (economics, sustainability, tourism, for example). Some examples of complex questions that a DSS user might want to ad-

*Corresponding author. E-mail: adrian.brasoveanu@modul.ac.at.

dress: Do financial crises have any effect on the tourist behavior? Do temperature increases in continental Europe change tourist behavior? Can the failure of some specific stocks predict a financial crisis?

Statistical data sources from multiple domains are increasingly available as linked data following the publication of the RDF Data Cube Vocabulary (QB) [1]. Visualization seems to be the de facto method for making sense of Linked Data (LD), and various approaches have been developed for navigating the data deluge (including for statistical LD), but less effort was dedicated to integrating visualizations into analytical platforms for answering complex questions, similar to the ones we have shown earlier. Fox and Hendler [14] argue that integration and reusability are the most important aspects on which visualization designers need to focus for succesfully controlling the current data deluge through visualizations.

This paper describes a methodology to integrate cross-domain statistical data sources by applying selected QB principles (observations and slicing, for example) to a Multiple Coordinated Views (MCV) dashboard. A tourism use case for cross-domain datasets from multiple sources including Eurostat and the World Bank is presented, and we discuss specific types of tasks that a visual dashboard helps to address.

The main contribution of the paper is threefold:

- a workflow and a set of visualization principles to be used for visualizing datasets in the RDF Data Cube vocabulary (Section 4);
- a collection of visualization scenarios that are useful when visualizing cross-domain data (Section 5);
- and a dashboard developed following these principles and scenarios. (Section 6).

The paper is organized as follows: the next Section offers a brief introduction to the QB vocabulary and a brief problem statement; Section 3 describes the current state-of-the-art in Semantic DSS and statistical LD Visualization; Section 4 describes the workflow and principles we have used to visualize statistical LD using different visual metaphors; Section 5 describes our tourism use case: the typical tasks that such a system needs to implement; Section 6 describes the design, implementation, and usage of a dashboard that meets the requirements of our use case; Section 7 contains the summary and conclusions.

[1] http://www.w3.org/TR/vocab-data-cube/

2. Background and Problem Statement

2.1. Background - RDF Data Cube Vocabulary

The RDF Data Cube Vocabulary is the current standard for publishing statistical data, and is a W3C Recommendation supported by industry and academia. QB has already gained acceptance by the community judging from the increasing number of statistical datasets published by using this vocabulary [2]. A further advantage of QB is that it is based on a cube model that is compatible with the SDMX standard (Statistical Data and Metadata Exchange) and designed to be general so that it enables the publishing of different types of multidimensional datasets.

The basic building blocks of the cube model are *measures, dimensions* and *attributes*, collectively referred to as *components*, and have the following roles:

- *Measure components* describe the things or phenomena that are observed or measured, for example, indicators such as height, weight or, in our tourism context, arrivals, bednights or capacity.
- *Dimension components* specify the variables that are important when defining an individual observation for a measurement. Examples of dimensions include time and space.
- *Attributes* help interpret the measured values by specifying the units of measurement, but also additional metadata such as the status of the observation (e.g., estimated, provisional).

Observations are the unit elements in a dataset and they represent a concrete measurement value for a set of concrete dimension values. They correspond to a value in a statistical database. When the value of a dimension is the same in a large number of observations (for example, the geographic location) it is convenient to group these into a slice. A dataset that contains observations grouped into slices across dimensions constitutes a cube. Each dataset is described by a Data Structure Document (DSD) that contains all the namespaces and components needed by that dataset.

2.2. Problem statement

Today's interconnected world exposes us all the time to various instabilities of nonperiodic flows sim-

[2] see for example the datasets listed at http://wiki.planet-data.eu/web/Datasets and the use cases described in http://www.w3.org/TR/2013/NOTE-vocab-data-cube-use-cases-20130801/

ilar to those described by Lorenz [35] and his peers (small changes of amplitudes lead to instabilities). If we are to design a DSS for domains that involve people (finance, tourism or culture, for example), we need to take into account the migration of people or their financial needs. This calls for the need to understand various financial and cultural profiles, because a person who grew up in such diverse places as Bucharest, Tallin, Vienna and London, might have a different set of aspirations and lifestyle than a person who lived all her life in Berlin. Such problems are easier to look at through the lens of statistics. In fact an immediate method to reduce the complexity derived from such phenomena is to use large collections of statistical data like those provided by the World Bank or Eurostat, which are now increasingly available as LD. These collections will help us to understand the macro perspectives, and can later be used to explain more complex phenomena.

By splitting the statistical data into cubes with maximum 3 dimensions, the QB vocabulary offers a simple structure that can be exploited to build this macro perspective, as this structure is shared by all datasets. Performing ontology alignment between any QB datasets is a problem that is usually complicated by several variables (lack of DSDs, failure of the SPARQL endpoint, mistakes in the data or DSD, the fact that QB is more like a guideline and the actual implementations can sometimes go in different directions, and so on). Also, if we really want to understand these macro trends, gathering the data will not suffice, therefore we will need to use visual methods to help during the decision-making process. This complicates the problem even further, as most visualizations are built for simple use cases. What if we want to display multiple coordinated visualizations built from a single query? Or what if we want to show both data analysis and visualizations in a single screen? Building visualizations is a time-consuming process, therefore we might want to reuse some of them. This leads to other challenges. What are the best design patterns for implementing reusable visualizations? Do existing interaction patterns from our visualizations need to be adapted for new datasets? Such questions lead us to the main problem we are interested in: *What are the principles, workflow and implementation design patterns that we need to follow in order to build a visual DSS that exploits cross-domain information?*

3. Related Work

As described in [1], Semantic Web (and Linked Data, by extension) and DSS can be viewed as application areas of Artifical Intelligence (AI), and in many cases the result of research in such an applied field is a system. However, if AI systems are to be effective, they need to use a variety of technologies in order to deliver their best results. In Semantic Web, for example, there is an increased wave of hybridization with Natural Language Processing (NLP), Machine Learning (ML), and Information Retrieval (IR), even the most popular systems like Watson subscribing to this trend [29,53]. Another possibility is to use Human-Computer Interaction (HCI) techniques like visualization to navigate the data flow.

3.1. Decision Support Systems

Semantic DSS. A survey of *Semantic DSS* is presented in [1]. Along with an overview of the systems, it presents a set of interviews with various research and industry partners. It identifies two main challenges for future Semantic DSS: a) the lack of flexible integration of information (most system do not integrate text, data and visualization well) and b) numerous issues related to the data analysis (cleaning, querying, aggregation, abstraction, etc) and scalability.

Tourism DSS. Our use case comes from the tourism domain, therefore we consider that taking a closer look at Tourism DSS is important. They have been implemented in various areas such as destination recommendation systems. Examples of previous research include studies about case-based travel recommendations [42], travel decision styles and destination recommendations [56], and creating adaptive recommender systems using neural networks [38]. Some well-known destination recommendation systems include DieToRecs [43], TourBO [15] and MobyRek [44]. Today travel recommendation systems are directed towards the consumer (i.e., the individual tourists) rather than the tourism managers therefore focusing on different content (i.e., destinations, touristic offers, events). BASTIS[3] and PATA (Pacific Asian Travel Association) mpower[4] integrate various types of data in one system, but unfortunately they come with pre-packaged visualizations that lack interactivity, and most of the data is hard-coded. The TourMIS [55] interface already supports

[3] http://bastis-tourism.info/
[4] http://mpower.pata.org/

a variety of decisions which would help Destination Marketing Organizations (DMO) to take appropriate actions (trends; performance indicators for cities); but it does not have cross-domain decision-making capabilities. Our work starts with the TourMIS data, as it will be seen in a later section of this paper (Section 6).

From the above we conclude that (1) the majority of tourism DSS are aimed at tourists rather than tourism managers; (2) while some tourism DSS systems aimed at managers include data from multiple data sources, this data is integrated in a hard-coded manner. We therefore see an important market niche for a new type of touristic dashboard targeted at managers and decision-makers, especially if we consider the ability to interactively integrate and visualize cross-domain data.

3.2. Linked Data Visualization

When it comes to LD visualization, we can distinguish two large domains: *ontology visualization (TBox visualizations)* and *instance data visualizations (ABox visualizations)* (in fact the visualization of the instances and the relations between them). Systems that offer both are also possible.

Ontology visualization. We consider the following surveys to offer a clear perspective on the evolution of the subject: a book and a survey about visualization in the early days of Semantic Web [17,33]; a survey of the role of ontologies in building user interfaces [40] and a more recent review focused on OWL visualizations [12]. The paper by Dudas [12] goes beyond a simple review, and not only examines the types of tasks needed in an ontology visualization system, but also analyzes how these tasks are supported in the current systems, and proposes an Ontology Visualization Recommender tool. While not included in these surveys, RDF based languages that allow us to visualize data or ontologies or both are now being used to visualize ontologies. RDFS/OWL Visualization Language (RVL) [41] was designed to create simple mappings between RDFS/OWL and D3.js [2] visualizations. It is a declarative language that allows creating visualizations from both the TBox and the ABox of a dataset. Another development is a visual language called VOWL2 [34] geared towards helping users visualize ontologies.

Instance data visualization. The landscape of instance data visualization is complex and contains many different types of visualization tools, as often LD is seen as just another type of data by the visualization designers. An early survey of the LD visualization techniques that predates the release of the RDF Data Cube Vocabulary can be found in Dadzie and Rowe's paper [9]. The survey presents the first coherent set of principles for visualizing LD, and proceeds to split the tools in two groups: text-based and LD browsers that offer visualization options. A survey of LD exploration systems [37] starts with a list of "exploratory search task characteristics" and links them to the features already implemented in LD browsers. The survey identifies three types of LD exploration systems: a) LD browsers; b) LD Recommenders; and c) LD based exploratory search systems. It then goes on to offer a systems timeline and a good summary of the best systems together with a comprehensive list of their IR and HCI features. Similar to the case of ontology visualization, there is also the possibility to use declarative LD for creating LD instance data visualizations, with RVL [41].

Since current LD systems tend to deliver data in all formats through their REST APIs, as we will see in the next subsection, we think there is a need to expand the principles of LD visualization towards particular types of datasets, as visualizations need to take into account the underlyining data structures.

3.3. Statistical Linked Data Visualization

Statistical LD visualization is a particular case of instance data visualization. Focusing on this type of visualizations, we can identify three large groups of *Statistical LD Visualizations*:

- tools and packages that offer *basic LD visualizations* (tables, charts, maps) of QB datasets, with or without aggregations;
- complex tools that integrate several visualizations into *dashboards using Multiple Coordinated Views (MCV)*;
- *LD platforms* that might contain visualizations

Basic visualizations and aggregations. The LOD2 project developed a Statistical Workbench that includes several components and reflects various phases of the statistical LOD consumption cycle, for example triplification (through CSV2DataCube), validation (through RDF Data Cube Validation tool) and visualization (with CubeViz)[13,48]). *CubeViz* [48] is an RDF Data Cube Browser which can be used to query both resources and observations from QB datasets, and display the results in the form of several classic chart types. The *OpenCube* toolkit offers another take on the statistical LOD lifecycle idea [28]. It includes

components for ETL (Grafter framework), a tool for data conversion (TARQL adaptation) and D2RQ extensions for publishing data cubes. For consuming the data they offer several tools: OpenCube Browser for table-based views, an R package for statistical analysis, a widget for slicing the data cubes, a catalogue management component and tool for interactive map-based visualizations of geographical data. As it can be seen, while the collection of use cases and tools from OpenCube is impressive, they are not integrated into a single package. While not necessarily a visualization tool, *Vital* [10] uses visualizations to help in the analysis and debuggging process for QB datasets publication. The automated *Visualization Wizard* described in [39] offers support for vocabulary mappings, considers the possible combinations of dimensions and measures for RDF Data Cubes, and offers a choice between several visualization packages (D3.js [2] and Google Charts). Another paper related to the same project [25] presents the *Linked Data Query Wizard*, a table-based approach to selecting query results from QB datasets, and uses classic chart types or mind maps to visualize the results. Ba-Lam Do [11] developed a visualization pipeline focused on creating *Linked Widgets* like lines, bars, pies, and especially maps, from QB datasets. He also identified two main problems for statistical LD visualizations: a) the challenge of analyzing and aligning multiple datasets due to the fact that most publishers use the QB vocabulary as a guideline and almost always come up with some changes to it; b) the challenge of creating tools for consuming statistical LD.

Dashboards built using Multiple Coordinated Views. All the visualization papers related to OLAP and data cubes could be included as related work here, but we chose to focuse more on the papers that present QB visualizations, or at least visualizations based on the Statistical Data and Metadata eXchange (SDMX) format, since this data can be represented as a Linked Data Cube. SDMX is the ISO standard format for statistical data representation currently used by large institution (IMF, the World Bank, Eurostat, and others). An early example of a geovisual analytics tool for regional data is the one built for visualizing OECD data by Jern and his team [27]. This example does not uses QB data as it was produced before the QB standard came out (2009), but the SDMX standard on which QB is based, and the visualizations from this dashboard (cloropleth map, scatter chart and parallel coordinates) are quite similar to the ones found in modern frameworks based on d3.js. Another example of LD visualization based on SDMX data is Hienert's dashboard [24], which allows you to add multiple visualizations on the same screen. Recently, the focus has moved towards QB visualizations. One of the first examples [45], is a follow up to the work described in [39] and [25], and it allows brushing over multiple coordinated visualizations. Sabol's paper analyzes two scenarios (search and analysis over LOD, analysis of scientifc publications), describes the workflow used to implement them, and the resulting visualizations in the extensions of the Visualization Wizard tool. Another tool that can be included in this category is LOD/VizSuite [52], but it only allows us to generate graph-based visualizations. Directly related to our use cases, the work of Benedikt Kämpgen and Andreas Harth, is focused on interogating multiple QB datasets via the OLAP4LD framework in [30] and [31]. Our own tool belongs to this category, and is based on the design philosophy presented in [49]. What makes our tool unique to the best of our knowledge, is the fact that it addresses well some of the challenges identified in the Semantic DSS study [1], as it will be seen in the following sections.

Linked Data Platforms [5]. The idea behind LDPs (Linked Data Platforms) is to deliver the data in various formats using REST APIs. Some platforms also allow to build full-featured interfaces that cand contain maps or pictures, typically using templating solutions like Velocity[6]. Elda[7], Carbon LDP[8], Apache Marmotta[9], Graphity[10], LDP4j [19] and Virtuoso[11] are several exponents of this trend. Many applications developed using following the LDP best practices[12], do include maps or other types of visualizations, as it can be seen from the following examples: *Bathing Water Quality*[13] or *Ordnance Survey*[14], both built with Elda. Another reason to include them as a separate type of visualization is the fact that many of these platforms are in fact used to publish QB datasets.

Some tools, while not directly related to QB datasets, might be useful when working with statistical data on

[5] http://www.w3.org/TR/ldp/
[6] http://velocity.apache.org/
[7] http://www.epimorphics.com/web/tools/elda.html
[8] https://carbonldp.com/
[9] http://marmotta.apache.org/
[10] https://github.com/Graphity/graphity-client
[11] http://virtuoso.openlinksw.com/
[12] https://dvcs.w3.org/hg/ldpwg/raw-file/default/ldp-bp/ldp-bp.html
[13] http://www.epimorphics.com/web/projects/bathing-water-quality
[14] http://data.ordnancesurvey.co.uk/

the web [26] presents a tool which allows visualizing provenance information, while [36] describes a tool for generating and viewing extended VoID descriptions of RDF datasets.

Most of today's visualization workflows are geared towards creating simple charts, and little effort (except for the MVC dashboards) is dedicated to complex analytic solutions that can answer the types of questions we are interested in. Without synchonizing multiple visualizations with the underlining data, and without combining multiple datasets, we think that it will be very hard if not impossible, to clearly present, in a clean user interface, complex use cases like those described in Kämpgen and Harth's work or at the beginning of this article.

4. Statistical Linked Data Visualization Principles

A number of formal models have appeared for describing *LD visualization workflows*. Typically, these models have been closely associated with prototypical implementations. Brunetti's [3] Linked Data Visualization Model (LVDM) is an extension of Chi's data state reference model [8] and consists of a series of transformation stages built on top of RDF and non-RDF data: a) *data transformation*; b) *visualization transformation*; c)*visual mapping transformation*. Helmich [23] uses Brunetti's model as implemented in Payola for visualizing the Czech LOD Cloud. De Vocht's [52] Visual Exploration Workflow is a pipeline and executable model for visualizing graphs that contains four types of views: a) overview groups, b) detailed groups referred to as narrowing views; c) coordinated views; and d) broadening views. Ba-Lam Do's Linked Widgets mashup platform is another example of a good visualization model. All these models and workflows resonate well with Schneiderman's Visual Information Seeking Mantra: *Overview first, zoom and filter, then details-on-demand*[50]. Shneiderman's taxonomy actually goes beyond this mantra and contains additional tasks: relate, history and extracts, as well as several specific visualization types. After almost two decades, this taxonomy is still one of the most popular for explaining the visualization process.

An extension of the task types taxonomy for interactive dynamic analysis can be found in Heer and Shneiderman [22]. The updated taxonomy contains twelve types of tasks split into three groups. *Data and view specification* (visualize, filter, sort, derive) tasks for exploring large datasets tend to focus on the selection of visual encodings rather than the actual visualization. For highlighting and coordinating interesting items, there is a category for *view manipulation* (select, navigate, coordinate, organize), which represents the core tasks in the original Information Seeking Mantra. Since today's visualizations are typically related to multiple datasets or articles, the last category of tasks is related to *process and provenance* (record, annotate, share, guide).

A list of the quantitative visualization types can also be found in [50], while a recent update can be found in Jeffrey Heer's visualization zoo [21]. An extensive treatment of the various reusable quantitative visualizations can be found in [54]. The book presents a grammar of graphics that allow us to build any 2D scientific visualization from a set of simple primitives like points, lines, scales or shapes. Recent visualization libraries built on top of D3 like like ggD3 [15], Vega[16] or NVD3[17] are following this philosophy.

Before explaining how to design visualizations following QB principles, we explain the workflow of statistical LD visualizations, as well as the tasks or visual metaphors involved.

4.1. Method and Workflow

Similarly to many other products visualizations are being created, used and replaced as part of a sequence of processes, often iterative in nature. This requires developers and users to follow a certain workflow. Since state-of-the-art systems put increased emphasis on automation and reuse, such a lifecycle can be expressed through a series of visualization pipelines ([11]; [28]). The abundance of different methodologies complicates the possibility to reuse existing statistical LD visualization workflows.

Building on best-practice examples reported in the literature, we propose a workflow that follows the logical sequence of developing statistical LD applications. We have closely followed this sequence when implementing the dashboard described in Section 6:

– **Requirements** - Application scenarios need to be well-understood to produce a good visualization. A scenario description should include the motivation and research questions to be explored, example data sources, and the type of visualizations ap-

[15] http://benjh33.github.io/ggd3/
[16] http://trifacta.github.io/vega/
[17] http://nvd3.org/

propriate to address the research questions. Special emphasis should be placed on visualization linking and the context of an application at an early stage.
- **Discovery and selection of indicators** - Running a sequence of SPARQL queries can yield an abundance of data, but to create real value the various dimensions of the datasets under consideration need to be analyzed, augmented or aggregated in order to fit particular visualization scenarios.
- **Ontology alignment** - There are a number of important steps that need to be taken into account when performing ontology alignment between QB datasets: broken or missing DSDs, failure of SPARQL endpoints, broken dumps. All these have to be included into alignment queries or scripts.
- **Indicator storage and retrieval** - Storage addresses the problem of failing SPARQL endpoints, and using effective indexing strategies in conjunction with established platforms such as ElasticSearch, Lucence or Sindice are essential when building IR applications.
- **Transformation** - This includes the specification of data wranglers [32] (scripts that transform data into formats suited for particular visualizations), queries or aggregations. Data items that are already indexed using a search server do not require data wrangling scripts, as the indexer already performs this mapping function. The transformation step can be seen as a first part of Heer and Shneiderman's data and view specification (filter, derive), even though derive tasks can also appear in subsequent steps [22].
- **Visualization** - Basic visualizations such as line charts, bar charts or pie charts tend to be reusable components. The focus on reusability resulted in visualization grammars, which can be considered the second part of Heer and Shneiderman's data and view specification (visualize, sort) [22].
- **Interaction** - An interaction layer (selections, zoom, pan, transitions, synchronization) is usually built on top of the visualization layer. This level corresponds to Heer and Shneiderman's view manipulation [22].
- **Reuse or Adapt** - Reuse can happen on multiple levels, from the indicators or indexes to specific charts or the entire platform. Reuse should be an integral part of the design process, parts of it corresponding to process and provenance in Heer and Shneiderman's taxonomy [22].

Any workflow for visualizing statistical LD includes both LD tasks (selection of indicators, ontology alignment, etc.) and visualization tasks (data wrangling, interaction, etc.). Due to the increased specialization of certain layers - e.g., alignment or interaction, visualizations represent collaborative processes. While visualization taxonomies are often straightforward to adapt to specific workflows, the LD processing differs from case to case, as statistical LD is still in the early stages of development.

4.2. Applying RDF Data Cube Principles

This section applies the following principles based on the RDF Data Cube design to the visualization of statistical LD: (i) Use of multiple coordinated views for linked visualizations, (ii) integration of data analysis and visualizations processes, (iii) visualization of slices instead of datasets, (iv) switching the fixed dimensions for slices with multiple fixed dimensions, (v) highlighting particular observations, and (vi) extracting and sharing visual knowledge. These principles do not modify those presented by Dadzie and Rowe [9], but rather extend them in the context of visualizing statistical LD.

The following list presents the principles we have followed when visualizing statistical LD.

- **Linked views for Statistical LD** - LD visualizations should reflect the linked nature of the data and support switching from one visualization to another when navigating the underlying datasets, or at the very least reflect changes across several visualizations on a single screen using multiple coordinated view technology. We call this principle the **Linked Visualizations for Linked Data** principle, and encourage it regardless of the nature of the linked datasets to be visualized. Linked visualizations are an obvious choice for statistical LD as statisticians tend to use multiple graphics to understand statistical phenomena.
- **Integration of data analysis and visualizations** - Since statistics GUIs like R also tend to integrate code, data and visualizations, we also recommend to *integrate the data analysis and the visualization taks*. Code should not be integrated, except if the GUI is dedicated to programmers. Supporting views that do not necessarily contain visualizations while displaying slices of datasets is a good

way to apply this principle - e.g., a list of top customers can be arranged after certain criteria, or a table to display the results of a statistical test.

– **Visualize slices instead datasets** - When visualizing particular datasets, one needs to take into account their structural characteristics. Since statistical LD datasets will rarely (if ever) be visualized in their entirety, systems require the ability to *visualize slices*. The RDF Data Cube Vocabulary identifies the dataset itself (qb:dataset), its structure (qb:structure) and dimensions (qb:dimension), as well as the actual measures (qb:measure) and observations (qb:observation). An observation about bednights occupied by German tourists in Prague from a tourism dataset, for example, will include dimensions such as market (Germany), destination (Prague) and time interval (January 2010), and measures such as the number of bednights. This corresponds to the structure of observations reported by statistics agencies and is equally suited for any type of experiment that tracks data over time (psychology, sociology, physics, etc). Visualizing slices instead of entire datasets in a specific context (together with texts or data, for example) also increases the value of the information presented to the user.
– **Flexible mechanisms for selecting slices** - Slices are collections of observations, in which at least one dimension remains fixed, approximating the way humans tend to query datasets, for example: *Identify all data about Austria's GDP between 2008 and 2014* (Austria represents the fixed dimension) or *Find all observations related to bookings by German tourists in Prague between 2008 and 2014* (Germany and Prague are fixed dimensions). In the second example, there is no way of telling if the user is actually interested in data related to the German clients, or to data related to people who visited Prague, or both. Therefore the best way to present the results is to take into account both dimensions and provide two separate views. Implementing *switching mechanisms for the fixed dimensions allows for flexibility in the choice of slices to be visualized.*
– **Highlighting particular observations** - The "Highlight links" principle from Dadzie and Rowe's work[9] needs to be extended to take into account the structure of the datasets. When using multidimensional datasets (e.g., tourists visiting a particular destination) we also need to *highlight specific (best, worst) observations*, not just the links. To differentiate these top observations, they could be aggregated by location and color-coded by performance indicators, for example.
– **Extract and share** - One of the main principles behind LD is its accessibility in multiple machine-readable formats. A quick way to do this through visualizations is to export data slices into various formats. Customizable image export functions support the dissemination of new research insights, for example, and reflect the last principle we propose, that of *extracting and sharing visual knowledge*.

The next section will introduce a tourism use case for statistical LD. It will outline the analysis of user requirements, show how to transform these requirements into visualization scenarios, and discuss how to implement these scenarios using the presented principles.

5. Decision Support Scenarios in Tourism

The strength of LD technology is that it simplifies combining information from various data sources by making explicit links between those entities that are the same (e.g., two cities) or explicitly stating the relation between similar things (e.g., stating that one statistical indicator is narrower than another). The ETIHQ dataset [47] is linked to various data sources and as such it allows integrating data from multiple statistical sources. Depending on the number of statistical data sources combined (e.g., TourMIS, World Bank) as well as the number of indicators visualised from these sources (e.g., bednights, arrivals) a range of practical decision support scenarios can be supported.

We have already seen that most of the current tools do not allow to create scenarios for visualizing linked models in a single screen, or to easily reuse scenarios by changing their input data. However, answering complex questions often requires combining multiple data sources (e.g., World Bank, Eurostat, etc) and multiple indicators from these sources (e.g., GDP, Tourist arrivals, etc). This requires the ability to specify not only the possible combinations of dimensions and measures within a visualisation, but also the granularity of the datasets (yearly, monthly), their provenance, or the various statistical tests that are needed to validate the statistical models.

Consider, for example the design of a tourism decision support system that relies on several datasets: a) ETIHQ - a new version of the TourMIS LD, a key

source of tourism statistical indicators in Europe published in terms of QB [46]; b) World Bank; c) Eurostat. TourMIS data contains 3 types of dimensions: destination (or city), market (countries where tourists come from), and time. The main datasets contains information about arrivals, bednights, capacities, points of interests or shopping indicators from the main European tourist destinations. Granularity is both monthly and yearly. A design pattern that is in the same time a powerful visual metaphor called multiple coordinated views is used to display multiple visualizations and synchronize them (as described in [49]).

5.1. Types of Visualization Scenario in Statistical Linked Data

In order to describe our use cases better, we have devised a theoretical framework (see Table 1) that takes into account the provenance of the indicators, and several possible scenario types (ST). These scenario types allow us to tell different stories, and mix the visualizations according to the hypothesis we want to check, but also with respect to data provenance.

The most common scenario type is **Scenario 1 (one indicator, one source - 1:1)** allows us to inspect one indicator from one data source, such as showing the TourMIS bednight indicator over a period of time, and is the simplest and most straightforward scenario that corresponds to the current functionality offered by TourMIS. Having a single indicator hardly restricts the visualization design space, as arrivals from different markets for the same destinations, can be shown via a large number of visual metaphors (line charts, bar charts, pie charts, arc diagrams, hive plots, etc). Different selections of the same indicator can be displayed on the same graph. By fixing destination, we can show values for different markets JPN (Japan), UK, GER (Germany) and answer simple questions (What are the top markets for certain cities?). By fixing market, we can show values for different destinations (UK arrivals to Vienna vs Linz vs Graz) with the goal of comparing destination performance. We can easily ask the same questions at country-level instead of city-level, by using the aggregation operators.

Visualization **Scenario 2 (two or more indicators, same source - n:1)** allow inspecting two (or more) indicators from the same source - for example, by display bednights and arrivals from the same market to a destination one could infer the percentage of the arriving tourists which sleep at hotels at that destination. Based on feedback from our tourism colleagues, this scenario is however rarely used in practice. We are typically interested in this type of scenario when we want to have a list of all indicators related to a certain topic from a single source: for example, we want to know which type of arrival indicators appear in TourMIS (arrivals inside the city, arrivals at city borders, arrivals at hotels, etc).

A type of scenario that can often send the wrong message for the user, but which can be interesting for a dataset publisher is **Scenario 3 (one indicator, multiple sources - 1:n)**. Inspecting values of the same indicator (e.g., arrivals) from two (or more) data sources is the general use case for this scenario, e.g., comparing arrival indicator values from TourMIS and the World Bank. When implementing such a scenario, it must be ensured that the indicator in the two data sources is measured in the same way, i.e., it has same (or comparable) meaning and it has same (comparable) semantics for its dimensions. While useful to verify the correlation of data between data sources, this scenario could lead to problematic cases by suggesting to users that the indicator data from one source is incorrect. This might not even be true, as in some cases it might be that just the data collection methodology is different, and at least when taking the LD version of the indicator it would be difficult to spot such cases. Therefore, our tourism colleagues advise against focusing on such scenarios.

Probably the most interesting cases are covered by **Scenario 4 (multiple indicators, multiple sources - n:n or m:n)**. Hypothesis for interdisciplinary researchers are typically addressed by these types of scenarios. Questions can sound like this: How are the arrivals from a certain market influenced by the GDP growth in a market country? Do CO_2 emissions in a destination city have any effect on the arrivals to that destination? Interesting correlations can be made at this level, also by varying the settings of an indicator: we can for example compare the performance of a city with that of a country, or the performance of a certain month versus the same year. Such crossdomain indicator comparisons, are, according to our tourism colleagues, the really interesting cases, not covered (or really difficult to cover) by traditional database-style systems and where LD technologies could provide a real benefit. When implementing such scenarios it is important that the two indicators are linked based on the value of one of their dimensions, that is the same or compatible (e.g., if one has cities and the other country data, city data from that country can be added up). Additionally, indicator value ranges should be the same,

Table 1

Overview and examples of decision support scenarios depending on the number of combined data sources and indicators.

Sources / Indicators	1 indicator	2 (+) indicators
1 source	**Scenario 1 (1:1)**: Inspect one indicator from one source – *e.g., how do the arrivals from UK and JP in Vienna compare?* – *e.g., where do more UK tourists arrive when comparing Vienna and Linz?*	**Scenario 2 (n:1)**: Inspect at least two indicators from the same source – *e.g., which percentage of tourists arriving in Vienna actually sleep there? (as a delta between arrivals and bednights)*
2 (+) sources	**Scenario 3 (1:n)**: Inspect one indicator from at least two sources – *e.g., How do arrivals to Vienna compare as recorded in TourMIS and World Bank?* – *e.g., Is GDP for a specific country (Austria) the same in Eurostat and World Bank?*	**Scenario 4 (n:n and n:m)**: Contrast at least two indicators from at least two data sources – *e.g., How does the GDP of a market country (e.g.,Japan) correlate with Arrivals/Bednights in one (or more) cities (e.g., Vienna vs. Amsterdam)*? – *e.g., How does tourism impact the environment of the host country?*

or compatible in the sense that higher granularity data can be obtained from lower granularity data by additions (e.g., month vs. year, city vs. country).

6. The ETIHQ Tourism Dashboard

The visual dashboard[18] (Figure 1) we created is to the best of our knowledge the first visual semantic DSS that uses multidomain knowledge in tourism. The current dashboard combines information from TourMIS, World Bank and EuroStat. Its design is based on the scenarios we have already discussed in the previous section. It currently allows decision makers to select and concurrently visualise tourism, economic and sustainability indicators, though the number of indicators can be extended to any number of domains for which we can find statistical LD. While TourMIS provides European tourism indicators, we select economics and sustainability indicators from the other two sources. Data from TourMIS/ETIHQ rarely overlaps with Eurostat or World Bank data, therefore scenarios that compare same indicator from multiple sources do not appear in this dashboard.

Our dashboard has two large components:

– An *indexer* package that represents the Linked Data components and produces an ElasticSearch index.
– A set of *reusable visualization components* that are linked together to form a dashboard.

[18] http://etihq.weblyzard.com

.

After discussing the design of the scenarios that were important for this dashboard, we will examine how each component implements the workflow from Section 4.1.

6.1. Requirements for Cross-domain Visualization Scenarios

In the previous section we have described a series of scenario types that are frequently met when creating statistical LD visualizations. Cross-domain scenarios (type 4 scenarios) were identified as being the most common scenarios. This type of scenarios can easily be displayed in multiple graphics. We present several scenarios of this type, together with the visualizations that our users were interested in.

A first scenario (2 sources, 3 indicators) explores the *links between finance and tourism*. A portal that queries the data from all the mentioned sources should allow users to get answers to the following questions: Is there a correlation between GDP fluctuation in Japan and arrival of tourists in Vienna? Do GDP fluctuations impact differently different cities? (e.g., if GDP decreases, do Japanese choose to visit Barcelona more and Vienna less?) A number of visual aids can support this analysis: *line chart* displays GDP growth for Japan, Arrivals in Vienna, Arrivals in Barcelona (data must be consistent from the point of view of granularity - e.g., do not mix yearly and monthly data); *geo map* presents arrivals of JPN tourists to all cities; a *table* shows all EU cities and the value of JP arrivals for the selected period and allows us to sort through each

column; selecting a city on the map or in the table will add that city to the line chart.

A second scenario studies the *links between environmental sustainability and tourism*. How does tourism impact the environment of the host country? Is there a correlation between the number of bednights in a country and the sustainability indicators (e.g. CO_2 emissions)? If we can see the volume of any sustainability indicator in combination with the arrivals and/or bednights then we can see the impact of tourism on the environment (especially when the user chooses a period of time, eg. between 2000-2014 then it will be more visible if the negative effects such as CO_2 emission is also increasing with the number of bednights and/or arrivals). We conclude that tourism has (or not) an impact on the environment of the country. Visual aids for supporting this analysis: *line chart* that displays arrivals and/or bednights in Austria and CO_2 emission rates in Austria; *geo map* - shows arrivals of UK tourists to all cities; *table* shows all EU countries and the CO_2 emission rates for all countries for the selected period; a *barchart* shows a CO_2 emissions indicator with different colors in order to warn us if they reached critical levels or not.

A third scenario explores *unemployment rate and tourist arrivals*. Does the unemployment rate in a source country impact the number of tourist arrivals coming from that source market to a target destination? Is there a reverse correlation between unemployment rate in United Kingdom and arrival of UK tourists in Paris? We would expect less people traveling from UK when unemployment increases. When the unemployment rate in United Kingdom increases, does the travel behaviour of tourists changes significantly because tourists choose less expensive destinations (Ljubljana instead of Prague)? In addition to presenting the results through a similar interface to the ones described for ST4.1 and ST4.2, we could add *flickering concentric circles* on top of the geo map that would signify hot destinations in times of crisis.

From this analysis we conclude that decision scenarios of type 4 (see Section 5.1) are (1) useful to support complex decision making processes, but are (2) currently difficult to achieve with state of the art database-style technologies (see Section 3.1) due to the high data integration effort that they require. Therefore, these scenarios will provide not only practical value for tourism managers, but will also allow benefiting from the strengths of LD and semantic technologies in the area of data integration based on semantic links.

6.2. The Linked Data Layers

In order to implement such scenarios one needs access to a lot of indicators from various data publishers. The tourism data we have used represents the dumps of a new version of TourMISLOD [46] called ETIHQ, which contain tourism data about arrivals, capacities, bednights, points of interest and shopping items in QB format. For Eurostat and World Bank data we have used dumps of economics and sustainability indicators published in the 270 Linked Dataspaces repositories[19] by Capadisli. Some details about the publishing process can be found in [4,5].

We have used several approaches for collecting the data for visualization. One approach was to use Federated SPARQL, but quite often it resulted in queryTimeouts. Another approach was to write SPARQL queries or bash scripts (a combination of cat and grep commands can give us all the URIs that respect a certain pattern, for example) and run them against the dumps collected from the three services. What we ended up using in practice was indexing the data using a search server (ElasticSearch) and create a Search API that gets the data from all the sources. The indexing service we created provides all the functionality for the three LD layers we envisioned.

Since the URIs from Eurostat and World Bank published in the 270 Linked Spaces are well-designed, the *discovery and selection of indicators* is not a complicated process as all that is needed is to have a vague idea about the name of the indicator you need (therefore to have a part of the name). If you already know the indicator name and URI, then you can directly provide the URI for the new datasets. In the first phase, the indexer will harvest all triples from that location that match your criteria (for example, only the data for indicators that correspond to real geographic entities, and no entities that were invented for statistics (like *Germany+France* or *EU-Germany*); or only data for the last 10 years).

A simple process of harvesting the triples that match certain criteria would not offer us enough information for a visualization. Some additional tasks that might need to be performed are usually those related to *ontology alignment*. One such example of alignment is the geospatial alignment done by our indexer: Geonames URIs are used instead of the names of the actual locations, as the real names of the location might suf-

[19] http://270a.info/

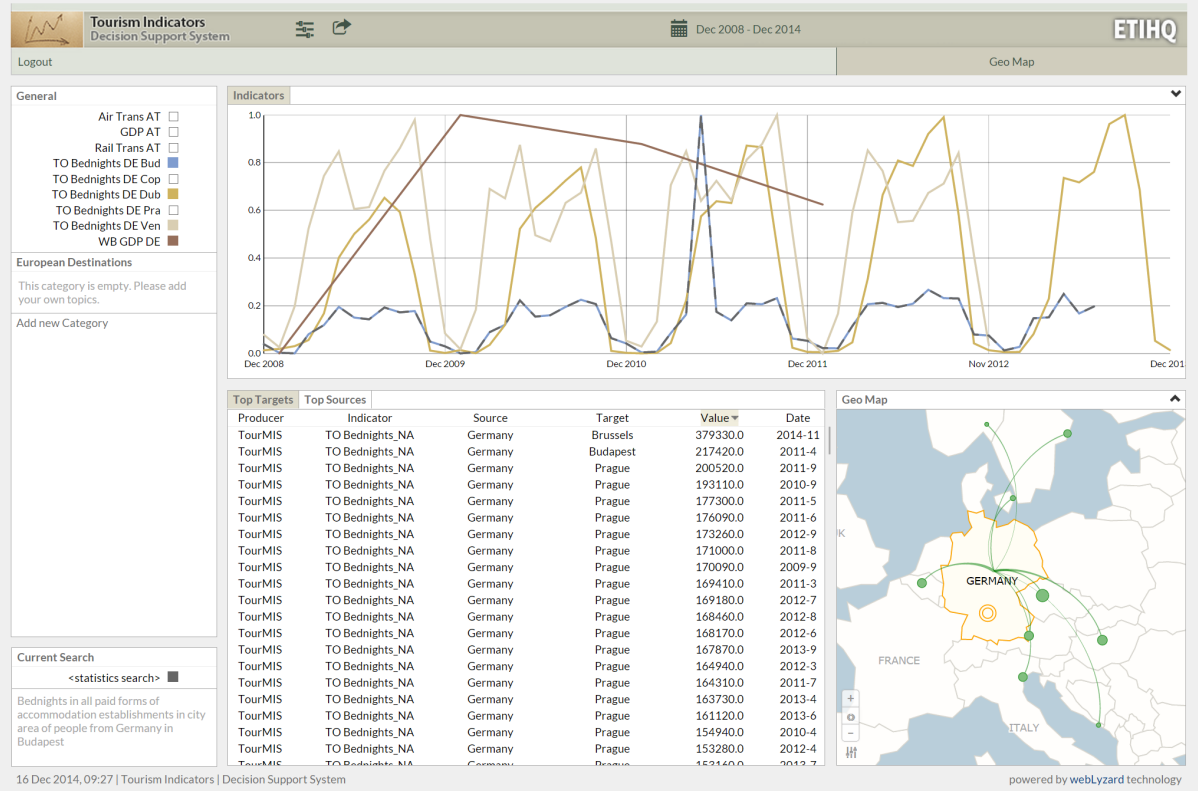

Fig. 1. The ETIHQ Dashboard. Bednights occupied by German tourists in various European destinations (Budapest, Dublin, Venice) plotted against the GDP growth of Germany.

fer from various issues like spelling mistakes, wrong encoding or even different name variants. Another example is the alignment of various units of measurements which was done using the DSDs (where they were available, else we took no units of measurements into consideration). We have not performed any alignment based on granularity of the temporal data (month, quarter, years), but instead used a convention: each observation corresponds to a data point in a graphic. The granularity information is added to each observation, and it can be used whenever it is needed (for complex aggregations at query time, for example).

When *indexing* the data, we have kept all the information (including the links) from the actual RDF dumps so that any observation or slice can be recreated if needed. The added information like granularity or geonames URI is only used for visualization purposes. It can be said that an indexer, in addition to the processing for the LD layers, also provides half of the functionality typically found on a transformation layer.

The functionality for all these layers (*selection of indicators, ontology alignment*) is included into a Python package. Somewhat similar components implemented as ElasticSearch river plugins[20] do not perform alignment, are focused on harvesting RDF as opposed to only RDF Data Cubes and have SPARQL queryTimeouts issues. We are considering releasing the code for the indexer package under an open source license, or opening the actual index for external users under different open and commercial licenses.

6.3. Cross-domain Visualization Layers

All the visualization layers are grouped in the actual dashboard product. The visualizations were designed taking into account the requirements presented in the previous sections (see Section 5 and Section 6.1).

The *transformation* layer contains the various queries and aggregations needed to feed the data into particular visualizations. There was no need for a data wrangling component as the data from the ElasticSearch index

[20] https://github.com/eea/eea.elasticsearch.river.rdf or https://github.com/jprante/elasticsearch-river-oai/

was already in the format needed by visualizations. As mentioned in the previous section, our indexer already performed half of the tasks usually found in this layer.

The *reusable visualizations* were written in JavaScript with jQuery and d3.js, following the conventions of the d3 reusable chart pattern [21]. All the visualizations are presented in a single-screen interface and are synced using the multiple coordinated views design pattern. For the rest of this section, we will discuss about the *visualization* and *interaction* layers. These two layers often mix, and we consider this situation normal, as often some types of interaction are easier to implement directly in a specific visualization, as opposed to exernal modules. It also helps us to present the workflow that we need to follow when constructing particular visualizations with our dashboard.

The current dashboard is targeted towards managers of DMO. This can be inferred directly from the scenarios we have considered in the previous section. A manager that wants to understand the influence of the financial crisis on the traveling behavior of German tourists, needs only to add some indicators to a chart, namely the variables he is interested in.

The manager can start with *adding an indicator* from TourMIS that shows the data slice representing the number of beds reserved by German tourists in Budapest. The definition of an indicator in the visual interface is a slice of data that covers the selected dates, and in which the market and the destination are fixed. Pushing the wheel button in the *General* pane (Figure 1) will uncover a menu where we will select *Add topic*. A topic corresponds to an indicator, that is a slice of the data in the respective interval (the time interval of interest must be selected in the upper-part of the interface) with market (source) and destination (target) as fixed dimensions. From the same menu we can *sort* the data from a chart alphabetically or by frequency.

It is recommended to create a *meaningful naming convention* for the topics / indicators, as shown in Figure 1, because the display space for menus will always be limited. Generally, we recommend that the names consist of the data source of the indicator (TO stands for TourMIS, ES for Eurostat and WB for World Bank), the name of the indicator (i.e., Bednights) and the dimension values that are chosen (in our example, these would be DE for Germany and Bud for Budapest). So, for this example indicator we provide the *TO Bednights DE Pra* name.

[21] http://bost.ocks.org/mike/chart/

Once named, a new indicator (or topic) is added on the right-hand panel of the portal, under the *General* heading. We then proceed to defining the topic. For this, hover over the new topic and press the wheel button that appears to its right. This action will replace the chart view in the top-middle pane of the interface with a dialog field that allows defining the topic, as shown in Figure 2. It enables selecting the data source (currently, World Bank, Eurostat, TourMIS), indicators (the indicators from the menu), markets and destinations (both can be cities or countries). A description of the selected indicator appears near the *Save* button. Once the relevant selections have been made, choose *Save*. This will close the dialog box.

For selecting the time interval, one needs to push the calendar button from the upper menu. This will open two date pickers that allow us to select the start and end dates we are interested in.

The *General* pane, the *Advanced Search* dialog, and the date selection mechanisms, allow the users to create most of the operations that sit on the *data and view specification* layers suggested by Heer and Shneiderman [22]. The *General* pane allows us to *filter* the indicators and *sort* them (through the wheel button's menu), and triggers the visualizations. By looking at the charts we can also *derive* new knowledge, this being the main purpose of designing a visual DSS.

As soon as the Advanced Search dialog box is closed, the data related to this topic is retrieved and visualized in the charts view (entitled *Indicators*). The first time a topic's data is visualized, the corresponding trend line is a dashed line. The current search can also be observed in the *Current Search* box, under the *General* menu.

The newly added topic also triggers various changes in the rest of the interface. The data displayed in the tables (middle pane) changes. This pane will create as many sub-panes as the number of dimensions for the visualized indicators. For our example, the TourMIS Bednights indicator has two dimensions, namely source and target, so two panes will be created corresponding to these dimensions (see the table in Figure 1). The *Targets* table, keeps the source value fixed (Germany, in our case) and varies the values for the Target cities, thus displaying the number of German tourists going to all European destinations. The table can be sorted based on the value field, thus allowing to quickly identify the most/least popular destination for Germans - it appears for example that Venice is a very popular destination for German tourists. Similarly, the *Source* table keeps the target fixed to Budapest, for

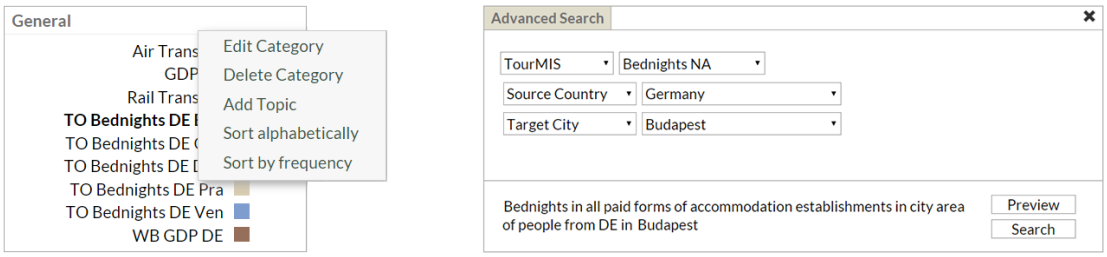

Fig. 2. Adding a topic: a) wheel menu; b) Advanced Search dialog for creating slices.

example, but varies the source markets, thus allowing detecting those tourist groups that go to Budapest the most/the least. World Bank and Eurostat indicators are from the economic and sustainability area, and therefore have a single dimension, that of the country/city of interest. In this case (as shown in the left side of Figure 3) a single table, called Targets, is created. The Targets table only contains data about the main markets for the indicator of interest.

A click on the pane name will trigger a change in the Geo Map (right pane of the interface), which displays the tabular data visually. The data for a particular market is summed up (from months to yearly data), and a visual representation of the connection between markets and destinations in the form of arrows is created (bigger arrows mean more tourists in the selected interval). The map from Figure 1, shows various destinations that were top choices for German tourists. For the Eurostat data (Air Transport indicator), the right side of Figure 3 (choropleth map), displays the markets using color coding (darker shades correspond to higher values), and the tooltips contain totals and averages of the selected indicator for the currently hovered country.

Since from the previous analysis Budapest does not necessarily stands out as a popular tourist destination for Germans (which is normal given the fact that it is not compared with anything), a new topic can be added that contains Bednights of German tourists to Budapest. This new topic can be added through the topic definition interface as explained before.

The previous steps allow exploring the behavior of German tourists in terms of their visitor volume to Budapest and also to other European cities. To understand whether this behaviour correlates with the economic situation in Germany, we can continue by selecting an economic indicator as a new topic. A good economic indicator is GDP Growth from World Bank (displayed as a brown line in the Figure 1). It might look like GDP growth starts from 0, in this picture, but in fact it has a negative value at the beginning of the interval, which is to be expected in times of crisis. Figure 1 super-imposes German GDP (from World Bank) as well as Bednights occupied by German tourists in Dublin, Venice and Budapest, as these indicators have been selected for visualization in the *General* pane (the color on the right side of a topic corresponds to the graph color on the chart - e.g., light blue for German Bednights to Budapest).

The resulting chart shows that there is a certain seasonality of the German visits in Budapest. The peak for each year is October (*Are Germans escaping from Oktoberfest?*). By inspecting German arrivals to several locations like Prague, Dublin, Venice and Budapest, it appears that German tourists seem to be influenced more by the seasonality of the business year (more visits during summer) than the crisis, as the patterns seem consistent from the end of 2008 to the end of 2014 and unaffected therefore by the slight GDP drop from 2009. Adding more destinations (Copenhagen, Dubrovnik, Venice) confirms our hypothesis of German tourist behavior being influenced by seasonality as opposed to GDP fluctuation.

These interconnected tables and charts correspond to Heer and Shneiderman's *view manipulation* logic [22]. We can *select* items from the tables and trigger new search with them as parameters, or we can *select* various observations from the line chart and display additional information in tooltips. The geomap allows people to see summaries about the various destinations visited by tourists from a certain country. Users can *organize* their workspace as they please, and are able to *coordinate* the views to explore the data in a meaningful way. Since everything happens on a single screen, *navigation* is reduced to several clicks in the various views.

Pushing the *Export* button, opens a side menu that allows us to select from two groups of options: *Chart Data* (XLS or CSV formats) and *Diagrams* (Line Chart, Geographic Map). They allow the DMO man-

Producer	Indicator	Target	Value	Date
Eurostat	ES Air Trans	Germany	180781584.0	2013
Eurostat	ES Air Trans	Germany	178591104.0	2012
Eurostat	ES Air Trans	Germany	175316080.0	2011
Eurostat	ES Air Trans	Germany	166130832.0	2010
Eurostat	ES Air Trans	Spain	165153232.0	2011
Eurostat	ES Air Trans	Spain	159771264.0	2012
Eurostat	ES Air Trans	Germany	158150128.0	2009
Eurostat	ES Air Trans	Spain	157731968.0	2013
Eurostat	ES Air Trans	Spain	153387008.0	2010
Eurostat	ES Air Trans	Spain	148318304.0	2009
Eurostat	ES Air Trans	France	138330832.0	2013
Eurostat	ES Air Trans	France	135005808.0	2012
Eurostat	ES Air Trans	France	131425312.0	2011
Eurostat	ES Air Trans	France	122887264.0	2010
Eurostat	ES Air Trans	France	117561568.0	2009
Eurostat	ES Air Trans	Italy	116226664.0	2011
Eurostat	ES Air Trans	Italy	116029392.0	2012
Eurostat	ES Air Trans	Italy	109064952.0	2010
Eurostat	ES Air Trans	Italy	101823760.0	2009
Eurostat	ES Air Trans	Netherlands	55680132.0	2012
Eurostat	ES Air Trans	Netherlands	53895216.0	2011
Eurostat	ES Air Trans	Netherlands	48616388.0	2010
Eurostat	ES Air Trans	Netherlands	46479064.0	2009
Eurostat	ES Air Trans	Switzerland	44216800.0	2013
Eurostat	ES Air Trans	Switzerland	43235844.0	2012
Eurostat	ES Air Trans	Switzerland	41439632.0	2011

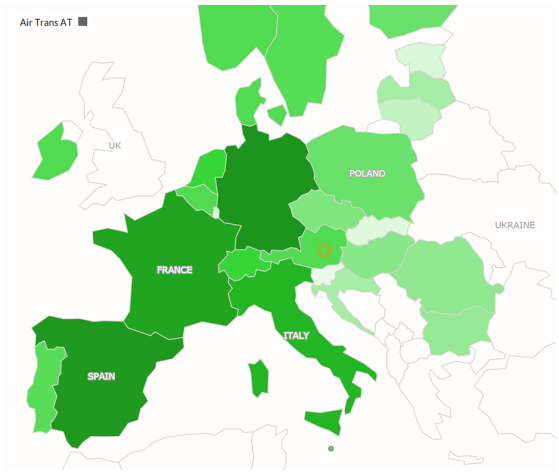

Fig. 3. Tabular and Geo-Map views of Eurostat data.

agers to *share* their work, create *guides* for their users or clients. The commercial implementation also allows to *record* and analyze the *Search History* (instead of the *Current Search* available in this version). These options represent our version of Heer and Shneiderman's [22] *process and provenance* functionality, and are one of our most popular features.

6.4. Reuse or Adapt

Current and future work is concerned with the last part of our workflow: *reuse and adapt*. Since all the components of our visualizations are reusable they can easily be added to any of our portals. The dashboard itself is a reusable component and can also be integrated into other products from the webLyzard ecosystem. In terms of scalability, it must be noted that each statistical LD dataset contains between several hundred thousands and 2 million observations, and that our solution supports an unlimited number of datasets.

These being said, we plan to focus our future work in three directions: a) incremental changes to the current tourism dashboard; b) integration with other portals and tools from the webLyzard ecosystem or from other ecosystems; c) improvements for on-the-fly data composition and visualization.

In the *incremental changes category*, we can include some features that are already planned for future releases. While people exposed to the prototype (especially managers from Tourism) have not requested it, we feel that it would be best to add several more data export formats (various RDF serializations, for example) and turn this dahboard into an even more complex multiple coordinated views based LDP hybrid (the first steps in this direction have already been done, as it can be seen in this article). Another small step is the integration of various LD sources. A good fit for the tourism portal would be the integration of smart cities data championed by Guimerans [18] and Celino [7], and the reuse of local governamental data proposed by De Vocht [51].

With respect to the *integration with other portals and tools* from the webLyzard ecosystem, we plan to permeate these statistical visualizations through several portals. In the Media Watch on Climate Change we plan to include these visualizations with a focus on climate change data. Since our portals also include news media, we will also be able to perform large scale news fact-checking, taking Tarasova's ideas [6] one step further. If there is interest in this area, we might integrate such multiple coordinate views LDP platforms with tools from other vendors.

On-the-fly data composition and visualization is now slowly turning into a hot topic. The idea to create new indicators by aggregating or simply adding the values from other indexes is presented in Gayo's work [16]. This opens the gate towards more complex analytical models with powerful prediction capabilities and will be explored in further publications.

7. Summary and conclusions

Decision-making processes can benefit a lot if they are supported by visual aids. A source of data for creating graphics for decision-making is statistical LD, as this type of data help us gain a macro perspective for many problems we face: collecting local governamen-

tal data, financial or health data, or even tourism data. Facilitated by the adoption of LD technologies, applications that seamlessly integrate and visualize statistical LD from multiple sources have just started to appear. The large scale integration of statistical LD technologies at semantic and syntactic level is still in its infancy, but better methods to align, link and visually explore datasets are needed.

In this paper we described advances to the state of the art in terms of (1) workflow and design principles for statistical LD visualizations from multiple sources; (2) visualization use cases for cross-domain statistical data; and (3) creating a visual DSS that implements the workflow, design principles, and the scenarios discussed in order to support cross-domain decision-making in tourism. Our solution integrates data analysis and visualization, but also comes of as a hybrid between multiple view coordinated solutions and LDP, and it can be easily reused or adapted. While it is not open-source, we are considering open-sourcing parts of it (the selection, discovery and alignment of indicators, for example).

We like to think that the innovation currently happening in the statistical LD space is just the beginning, and that in few years, these improvements will lead to complex decision-making processes and tools. In fact such data sources and the tools that visualize them, represent the best long-term investment we can make if we want to understand systemic issues that plague our society, as it is currently hard to argue against "the unreasonable effectiveness of data" [20]. A decision-making tool is only as good as its data, and as we have seen there are real insights to be gained.

Of course, while we have not focused on it in this article, the validity of the data is important. We have used data produced by us and a third part we were able to trust, but if large scale LD analytics solutions will succeed in the next few years, people need to be able to trust the results they see on their screen. The very open nature of LD can be the main issue here, as we feel there should be an independent international organisation that certifies data publishers not only with respect to the quality of the published data, but also with respect to following security standards.

The workflow and principles we presented can also be applied to other types of data: scientific data (which is of course statistical in nature, though not always published with QB standards), personal data (a CV or a FOAF profile could highlight important slices from a person's life), or historical data (the trends of various historical periods could be grouped together and the dominant ones highlighted). At this stage, perhaps each community must develop its own principles and a later comittee or survey will merge all of these into a set of principles for visualizing any kind of LD.

Acknowledgements

The work presented in this paper is partly funded by the PlanetData project (FP7:ICT - 2009.3.4, 257641). We thank Prof. Karl Wöber and Dr. Irem Onder for their guidance on tourism related issues. We also like to thank Dr. Alistair Jones, Daniel Kropshofer, Ruslan Kamolov and Walter Rafelsberger for advices and developing some of the visualizations.

References

[1] E. Blomqvist. The use of semantic web technologies for decision support - a survey. *Semantic Web*, 5(3):177–201, 2014.

[2] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Trans. Vis. Comput. Graph.*, 17(12):2301–2309, 2011.

[3] J. M. Brunetti, S. Auer, R. García, J. Klímek, and M. Nečaský. Formal linked data visualization model. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services*, page 309. ACM, 2013.

[4] S. Capadisli, S. Auer, and A.-C. Ngonga Ngomo. Linked sdmx data. *Semantic Web*, 2013.

[5] S. Capadisli, S. Auer, and R. Riedl. Linked statistical data analysis. *Semantic Web Challenge*, 2013.

[6] I. Celino and G. R. Calegari. Geo-statistical exploration of milano datasets. In S. C. et al, editor, *Second International Workshop for Semantic Statistics SemStats 2014*. CEUR-WS, 2014.

[7] I. Celino and A. Carenini. Towards a semantic city service ecosystem. In T. Omitola, J. G. Breslin, and P. M. Barnaghi, editors, *Proceedings of the Fifth Workshop on Semantics for Smarter Cities a Workshop at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014.*, volume 1280 of *CEUR Workshop Proceedings*, pages 3–8. CEUR-WS.org, 2014.

[8] E. H. Chi. A taxonomy of visualization techniques using the data state reference model. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pages 69âĂŞ–75. IEEE, 2000.

[9] A. Dadzie and M. Rowe. Approaches to visualising linked data: A survey. *Semantic Web*, 2(2):89–124, 2011.

[10] E. Daga, M. d'Aquin, A. Gangemi, and E. Motta. Early analysis and debugging of linked open data cubes. In S. C. et al, editor, *Second International Workshop for Semantic Statistics SemStats 2014*. CEUR-WS, 2014.

[11] B. Do, T. Trinh, P. Wetz, A. Anjomshoaa, E. Kiesling, and A. M. Tjoa. Widget-based exploration of linked statistical data spaces. In M. Helfert, A. Holzinger, O. Belo, and C. Francalanci, editors, *DATA 2014 - Proceedings of 3rd International*

Conference on Data Management Technologies and Applications, Vienna, Austria, 29-31 August, 2014. SciTePress, 2014.

[12] M. Dudás, O. Zamazal, and V. Svátek. Roadmapping and navigating in the ontology visualization landscape. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, volume 8876 of *Lecture Notes in Computer Science*, pages 137–152. Springer, 2014.

[13] I. Ermilov, M. Martin, J. Lehmann, and S. Auer. Linked open data statistics: Collection and exploitation. In P. Klinov and D. Mouromtsev, editors, *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings*, volume 394 of *Communications in Computer and Information Science*, pages 242–249. Springer, 2013.

[14] P. Fox and J. Hendler. Changing the equation on scientific data visualization. *Science(Washington)*, 331(6018):705–708, 2011.

[15] T. Franke. Tourbo: A prototype of a regional tourism advising system in germany. In D. R. Fesenmaier and K. W. Wöber, editors, *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 257–280. CABI, 2006.

[16] J. E. L. Gayo, H. Farhan, J. C. Fernández, and J. M. Á. Rodráguez. Representing verifiable statistical index computations as linked data. In S. C. et al, editor, *Second International Workshop for Semantic Statistics SemStats 2014*. CEUR-WS, 2014.

[17] V. Geroimenko and C. Chen, editors. *Visualizing the Semantic Web*. Springer, 2002.

[18] A. G. Guimerans, B. Villazón-Terrazas, and J. M. Goóez-Pérez. A linked data lifecycle for spanish smart cities. In T. Omitola, J. G. Breslin, and P. M. Barnaghi, editors, *Proceedings of the Fifth Workshop on Semantics for Smarter Cities a Workshop at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014.*, volume 1280 of *CEUR Workshop Proceedings*, pages 9–14. CEUR-WS.org, 2014.

[19] M. E. Gutiérrez, N. Mihindukulasooriya, and R. García-Castro. Ldp4j: A framework for the development of interoperable read-write linked data applications. In R. Verborgh and E. Mannens, editors, *Proceedings of the ISWC Developers Workshop 2014, co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014.*, volume 1268 of *CEUR Workshop Proceedings*, pages 61–66. CEUR-WS.org, 2014.

[20] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24(2):8–12, 2009.

[21] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Commununications of the ACM*, 53(6):59–67, 2010.

[22] J. Heer and B. Shneiderman. Interactive dynamics for visual analysis. *Communications of the ACM*, 55:45–54, 2012.

[23] J. Helmich, J. Klímek, and M. Necaský. Visualizing RDF data cubes using the linked data visualization model. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, volume 8798 of *Lecture Notes in Computer Science*, pages 368–373. Springer, 2014.

[24] D. Hienert, B. Zapilko, P. Schaer, and B. Mathiak. Web-based multi-view visualizations for aggregated statistics. *CoRR*, abs/1110.3126, 2011.

[25] P. Hoefler, M. Granitzer, E. Veas, and C. Seifert. Linked data query wizard: A novel interface for accessing SPARQL endpoints.

[26] R. Hoekstra and P. T. Groth. PROV-o-viz-understanding the role of activities in provenance. 2014.

[27] M. Jern. Collaborative web-enabled geoanalytics applied to OECD regional data. In Y. Luo, editor, *Cooperative Design, Visualization, and Engineering, 6th International Conference, CDVE 2009, Luxembourg, Luxembourg, September 20-23, 2009. Proceedings*, volume 5738 of *Lecture Notes in Computer Science*, pages 32–43. Springer, 2009.

[28] E. Kalampokis, A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris, and K. Tarabanis. Creating and utilizing linked open statistical data for the development of advanced analytics services. In S. C. et al, editor, *Second International Workshop for Semantic Statistics SemStats 2014*. CEUR-WS, 2014.

[29] A. Kalyanpur, B. Boguraev, S. Patwardhan, J. W. Murdock, A. Lally, C. Welty, J. M. Prager, B. Coppola, A. Fokoue-Nkoutche, L. Zhang, Y. Pan, and Z. Qiu. Structured data and inference in deepqa. *IBM Journal of Research and Development*, 56(3):10, 2012.

[30] B. Kämpgen and A. Harth. OLAP4LD - A framework for building analysis applications over governmental statistics. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, volume 8798 of *Lecture Notes in Computer Science*, pages 389–394. Springer, 2014.

[31] B. Kämpgen, S. Stadtmüller, and A. Harth. Querying the global cube: Integration of multidimensional datasets from the web. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, volume 8876 of *Lecture Notes in Computer Science*, pages 250–265. Springer, 2014.

[32] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3363–3372. ACM, 2011.

[33] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. G. Giannopoulou. Ontology visualization methods - a survey. *ACM Comput. Surv.*, 39(4), 2007.

[34] S. Lohmann, S. Negru, F. Haag, and T. Ertl. VOWL 2: User-oriented visualization of ontologies. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, volume 8876 of *Lecture Notes in Computer Science*, pages 266–281. Springer.

[35] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[36] E. Mäkelä. Aether - generating and viewing extended void statistical descriptions of RDF datasets. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, and A. Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, volume 8798 of *Lecture*

Notes in Computer Science, pages 429–433. Springer, 2014.

[37] N. Marie and F. L. Gandon. Survey of linked data based exploration systems. In D. Thakker, D. Schwabe, K. Kozaki, R. Garcia, C. Dijkshoorn, and R. Mizoguchi, editors, *Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014*, volume 1279 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[38] J. A. Mazanec. Building adaptive systems: A neural net approach. In D. R. Fesenmaier and K. W. Wöber, editors, *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 137–170. CABI, 2006.

[39] B. Mutlu, P. Höfler, V. Sabol, G. Tschinkel, and M. Granitzer. Automated visualization support for linked research data. In S. Lohmann, editor, *Proceedings of the I-SEMANTICS 2013 Posters & Demonstrations Track, Graz, Austria, September 4-6, 2013*, volume 1026 of *CEUR Workshop Proceedings*, pages 40–44. CEUR-WS.org, 2013.

[40] H. Paulheim and F. Probst. Ontology-enhanced user interfaces: A survey. *Int. J. Semantic Web Inf. Syst.*, 6(2):36–59, 2010.

[41] J. Polowinski. Towards RVL: a declarative language for visualizing RDFS/OWL data. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, page 38. ACM, 2013.

[42] F. Rici, D. Cavada, N. Mirzadeh, and A. Venturini. Case-based travel recommendations. In D. R. Fesenmaier and K. W. Wöber, editors, *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 67–93. CABI, 2006.

[43] F. Rici, D. R. Fesenmaier, N. Mirzadeh, H. Rumetshofer, E. Schaumlechner, A. Venturini, K. W. Wöber, and A. H. Zins. Dietorecs: A case-based travel advisory system. In D. R. Fesenmaier and K. W. Wöber, editors, *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 227–239. CABI, 2006.

[44] F. Rici and Q. N. Nguyen. Mobyrek: A conversational recommender system for on-the-move travellers. In D. R. Fesenmaier and K. W. Wöber, editors, *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 281–294. CABI, 2006.

[45] V. Sabol, G. Tschinkel, E. E. Veas, P. Höfler, B. Mutlu, and M. Granitzer. Discovery and visual analysis of linked data for humans. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. T. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, editors, *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, volume 8796 of *Lecture Notes in Computer Science*, pages 309–324. Springer, 2014.

[46] M. Sabou, I. Arsal, and A. M. P. Braşoveanu. Tourmislod: A tourism linked data set. *Semantic Web*, 4(3):271–276, 2013.

[47] M. Sabou, A. M. P. Braşoveanu, and I. Arsal. Linked data for cross-domain decision-making in tourism. 2015.

[48] P. E. Salas, M. Martin, F. M. D. Mota, K. Breitman, S. Auer, and M. A. Casanova. Publishing statistical data on the web. In *Proceedings of 6th International IEEE Conference on Semantic Computing*, IEEE 2012, Palermo, Italy, 2012. IEEE.

[49] A. Scharl, A. Hubmann-Haidvogel, A. Weichselbraun, H.-P. Lang, and M. Sabou. Media watch on climate change–visual analytics for aggregating and managing environmental knowledge from online sources. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 955–964. IEEE, 2013.

[50] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[51] L. D. Vocht, M. V. Compernolle, A. Dimou, P. Colpaert, R. Verborgh, E. Mannens, P. Mechant, and R. V. de Walle. Converging on semantics to ensure local government data reuse. In T. Omitola, J. G. Breslin, and P. M. Barnaghi, editors, *Proceedings of the Fifth Workshop on Semantics for Smarter Cities a Workshop at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014.*, volume 1280 of *CEUR Workshop Proceedings*, pages 47–52. CEUR-WS.org, 2014.

[52] L. D. Vocht, A. Dimou, J. Breuer, M. V. Compernolle, R. Verborgh, E. Mannens, P. Mechant, and R. V. de Walle. A visual exploration workflow as enabler for the exploitation of linked open data. In D. Thakker, D. Schwabe, K. Kozaki, R. Garcia, C. Dijkshoorn, and R. Mizoguchi, editors, *Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014*, volume 1279 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[53] C. Welty. Semantic web and best practice in watson. In S. Coppens, K. Hammar, M. Knuth, M. Neumann, D. Ritze, H. Sack, and M. V. Sande, editors, *Proceedings of the Workshop on Semantic Web Enterprise Adoption and Best Practice*, volume 1106 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[54] L. Wilkinson. *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

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

[56] A. Zins and K. Grabler. Destination recommendations based on travel decision styles. In D. R. Fesenmaier and K. W. Wöber, editors, *Destination Recommendation Systems: Behavioural Foundations and Applications*, pages 94–120. CABI, 2006.