Semantic Web 1 (2016) 1–25
IOS Press

Visualizing Statistical Linked Knowledge for Decision Support

Editor(s): Aba-Sah Dadzie, The Open University, UK; Emmanuel Pietriga, INRIA France & INRIA Chile
Solicited review(s): Luc Girardin, ETH Zurich, Switzerland; Bernhard Schandl, University of Vienna, Austria; Emmanuel Pietriga, INRIA France & INRIA Chile

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

[a] *Department of New Media Technology, MODUL University Vienna, Am Kahlenberg 1, 1190 Vienna, Austria*
E-mail: {adrian.brasoveanu, alexander.hubmann, daniel.fischl}@modul.ac.at
[b] *Christian Doppler Laboratory for Software Engineering Integration for Flexible Automation Systems, Vienna University of Technology, Favoritenstrasse 9-11, 1040 Vienna, Austria*
E-mail: marta.sabou@ifs.tuwien.ac.at
[c] *webLyzard technology gmbh, Puechlgasse 2/44, 1190 Vienna, Austria*
E-mail: scharl@weblyzard.com

Abstract. In a global and interconnected economy, decision makers often need to consider information from various domains. A tourism destination manager, for example, has to correlate tourist behavior with financial and environmental indicators to allocate funds for strategic long-term investments. Statistical data underpins a broad range of such cross-domain decision tasks. A variety of statistical datasets are available as Linked Open Data, often incorporated into visual analytics solutions to support decision making. What are the principles, architectures, workflows and implementation design patterns that should be followed for building such visual cross-domain decision support systems. This article introduces a methodology to integrate and visualize cross-domain statistical data sources by applying selected RDF Data Cube (QB) principles. A visual dashboard built according to this methodology is presented and evaluated in the context of two use cases in the tourism and telecommunications domains.

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

1. Introduction

Decision Support Systems (DSS) are typically customized for specific decisions in a given domain. In a global economy, external events such as financial crises or climate change - through observable consequences like bankruptcies or hurricanes - can render such domain-specific solutions obsolete. Building comprehensive monitoring systems into DSS tools is a potential solution, but one that can be prohibitively expensive and out-of-reach for smaller companies and research groups. One way to build DSS tools that leverage such cross-domain information is to analyze aggregated representations of events in the form of statistical data. Such an integration helps answer complex questions that require cross-domain data. Drawing on economic and sustainability indicators in conjunction with behavioral data from tourism research, for example, allows answering complex questions such as the following: Do financial crises affect tourist behavior? Do temperature increases in continental Europe change the annual distribution of arrivals? Can the failure of specific stocks - e.g., large tour operators or hotel stocks - predict a sector-wide crisis? Similar questions arise in other sectors as well. A telecommunications analysts, for example, might want to investi-

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

gate longitudinal data to explain unusual peaks in the number of calls or text messages sent, or better understand how geopolitical trends (migration, aging population, etc.) influence call data patterns.

Statistical data sources from multiple domains are increasingly available as linked (open) 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 [11], but less effort was dedicated to integrating visualizations into analytical platforms for answering complex questions, similar to the ones we have discussed earlier. Fox and Hendler [16] argue that *integration* and *reusability* are the most important aspects on which visualization designers need to focus for successfully controlling the current data deluge through visualizations. The success of tools like LOD, QB or D3 [4] has greatly simplified data publishing and visualization, but the problems described by Fox and Hendler still persist due to a combination of factors: i) the standards are often taken as guidelines and there is a lot of improvisation when publishing datasets; ii) integration of SLD is a complex field [56]; iii) projects are mostly focused on creating individual visualizations (line charts, bar charts, etc) instead than frameworks for integrating multiple types of visualizations; iv) in the context of Big Data, scalability has to be taken into account when designing new systems right from the start.

This article describes the methodology that was used to remove some of these gaps and to integrate cross-domain statistical data sources into a visual dashboard that supports a multiple coordinated view approach. A first prototype considered specific project requirements in conjunction with recommendations from Dadzie and Rowe [11] and QB (concepts such as *observations* and *slicing*). We have then extracted a set of principles and workflows for integrating and visualizing heterogeneous data sources that we later applied to various use cases (e.g., tourism, telecommunications, etc). We iteratively continued to improve and deploy new versions of this technology. The current article is focused on the first two generations.

We present use cases from the tourism and telecommunications domains based on cross-domain datasets from multiple sources including Eurostat[2] and the World Bank[3], and discuss specific types of tasks that the visual dashboard helps address. The article's main contributions include:

- a set of workflows and visualization principles usable for visualizing datasets in the RDF Data Cube vocabulary (Section 4);
- a collection of visualization scenarios that are useful for multiple use cases (Section 5);
- visual dashboards developed following these principles and scenarios (Section 6 and Section 8).

The remainder of this article is structured as follows: Section 2 offers an introduction to the QB vocabulary and formulates the problem statement; Section 3 describes the current state of the art in statistical LD Visualization; Section 4 describes the principles, architecture and workflows we propose to visualize statistical LD using different visual metaphors; Section 5 describes use cases from the tourism and telecommunication domains and how they guided the development of visual tools; Section 6 describes the design, implementation, and usage of a tourism dashboard in line with the use case requirements, which is evaluated in Section 7. The telecommunications dashboard presented in Section 8 builds on recommendations derived from this evaluation. Section 9 summarizes the lessons we learned and outlines future research avenues.

2. Background and Problem Statement

2.1. Background - RDF Data Cube Vocabulary

The RDF Data Cube Vocabulary is a W3C Recommendation for publishing statistical data, supported by industry and academia as evidenced by the increasing number of datasets published using this vocabulary; e.g., the PlanetData datasets[4] or the W3C use cases.[5] A further advantage of QB is that it is based on a cube model that is compatible with the *Statistical Data and Metadata Exchange* (SDMX) standard 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:

[1] www.w3.org/tr/vocab-data-cube
[2] eurostat.linked-statistics.org
[3] worldbank.270a.info/about.html
[4] wiki.planet-data.eu/web/datasets
[5] www.w3.org/tr/vocab-data-cube-use-cases

- *Measure components* describe the things or phenomena that are observed or measured (e.g., height, weight, arrivals, bed nights, capacity, number of mobile phone calls).
- *Dimension components* specify the variables that are important when defining an individual observation for a measurement (e.g., 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., observed, estimated).

These basic building blocks are then combined into more complex structures such as *slices* and *datasets*:

- *Observations* are the atomic data units that represent a concrete measured value for a set of concrete dimension values. Observations correspond to the values from statistical databases. Sometimes observations can also contain multiple measurements related to the same dimensions.
- *Slices* are groups of observations with several dimensions fixed (e.g., the arrivals of German tourists in Budapest between 2007 and 2013 has only one variable dimension: time).
- *Datasets* are collections of observations with the same dimensions and measures. Datasets that contain observations grouped into slices across dimensions constitute a *cube*.
- A *Data Structure Document (DSDs)* describes a dataset and contains all the required namespaces and components.
- *Code lists* or *dictionaries* describe the list of entities that are repeated through all datasets of a publisher (e.g., countries, units of measurements). They can also be used to describe complex hierarchies (geopolitical, ISO classification, etc.), and are often described using the SKOS vocabulary.

2.2. Problem Statement

Global economies expose us to various instabilities of non-periodic flows similar to those described by Lorenz [34]. Most domains reflect aggregated patterns of human behavior (finance, telecommunications, tourism, culture, etc.), where small changes of amplitudes can lead to instabilities. To design a DSS for such dynamic domains, one needs to understand financial and cultural profiles (migration patterns, financial needs, etc.). Such problems are easier to investigate 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 such as those provided by the World Bank or Eurostat, which are now increasingly available as LD. Such collections help understand macroscopic effects when investigating complex economic, environmental or social phenomena.

By splitting statistical data into cubes of up to three dimensions, the QB vocabulary offers a simple and flexible structure to represent such macroscopic effects. Performing ontology alignment between any QB datasets is a problem that is usually complicated by a number of factors - lack of DSDs, failure of SPARQL endpoints, errors in the data or DSD, deviations between QB guidelines and actual implementations, etc. Simply gathering a lot of data will not suffice to understand macro trends, however, and visual methods can help reduce data complexities during the decision-making process. Building a visual DSS is the first step towards a full-fledged DSS system, but the output of the visualizations (e.g., correlations, patterns) does not necessarily need to be translated into new knowledge (e.g., by creating new annotations or datasets with these correlations). Even without automatic interpretation of the results, this still complicates the problem, as most visualizations are built for simple use cases. What methodology needs to be followed to display multiple coordinated visualizations built from a single query? What are good methods to show both numeric results of analytic processes and the corresponding visualizations in a unified view?

Building visualizations is a time-consuming process, and the desired ability to reuse them poses a number of challenges. What are the best design patterns for implementing reusable visualizations? Do existing interaction patterns of existing visualizations need to be adapted for new datasets? These questions lead to the main research problem investigated in this article: *What are the principles, architectures, workflows and implementation design patterns needed to build a visual DSS that exploits cross-domain information?*

3. Related Work

A survey of *Semantic DSS* [2] contains an overview of the systems and 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 systems do not integrate text, data and visualization well) and b) numerous issues related to the data analysis (cleaning,

querying, aggregation, abstraction, etc) and scalability. Semantic Web (and Linked Data, by extension) and DSS can be viewed as application areas of Artificial Intelligence (AI), and in many cases the result of research in such an applied field is a system. However, effective AI systems need to use a variety of technologies 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 such as Watson subscribing to this trend [27,53]. Another possibility is to use Human-Computer Interaction (HCI) techniques such as visualization to navigate the data flow. The remainder of this section is focused on the visualization of Linked Data as a major means of sense-making.

3.1. Linked Data Visualization

We can distinguish two large domains of LD visualization: *ontology visualization (TBox visualizations)* and *instance data visualizations (ABox visualizations*). Systems that offer both are also possible. In both cases, the main goal of the visualizations is to help understand the relations between the various ontology classes or instances. We also discuss several Linked Data Visualization Models.

Ontology Visualization. The evolution of ontology visualization can be traced through several surveys about the early days of Semantic Web visualization [18,30]; the role of ontologies in building user interfaces [38] and OWL visualizations [14]. The paper by Dudas [14] also examines the types of tasks needed in an ontology visualization system, how these tasks are supported in the current systems, and showcases an Ontology Visualization Recommender tool. RDF based languages that allow visualizing ontologies and data are now being used to visualize ontologies. RDFS/OWL Visualization Language (RVL) [39] was designed to create simple mappings between RDFS/OWL and D3.js [4] 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 [33] geared towards helping users visualize ontologies.

Instance Data Visualization. The early survey of LD visualization techniques from Dadzie and Rowe [11] predates the release of the RDF Data Cube Vocabulary. It contains the first coherent set of principles for visualizing LD, and divides existing visualization tools into two groups: text-based and LD browsers that offer visualization options. A later survey of LD exploration systems [35] starts with a list of search task characteristics and links them to features already implemented in LD browsers. The survey identifies three types of LD exploration systems: LD browsers; LD Recommenders; and LD-based exploratory search systems. It offers a summary of best-practice systems including their IR and HCI features. Similar to the case of ontology visualization, there is the possibility to use declarative LD for creating LD instance data visualizations with RVL [39].

Linked Data Visualization Models. A number of formal models describe *LD visualization workflows*, some of them also being associated with prototype implementations. De Vocht's [50] Visual Exploration Workflow is an executable model for visualizing graphs that contains four types of views (overview groups, narrowing views, coordinated views and broadening views). Brunetti's [7] Linked Data Visualization Model (LVDM) extends Chi's data state reference model [10] 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 [22] implements this model in Payola for visualizing the Czech LOD cloud. Ba-Lam Do's Linked Widgets platform [13] is a pipeline for creating mashups.

All these models and workflows resonate well with Schneiderman's Visual Information Seeking Mantra: *Overview first, zoom and filter, then details-on-demand* [48]. Shneiderman's taxonomy actually goes beyond this mantra and contains additional tasks: relate, history and extract, as well as a list of the quantitative visualization types. A recent list of visualization types can be found in Heer's visualization zoo [20], while an extension of the task types taxonomy for interactive dynamic analysis can be found in Heer and Shneiderman [21]. The updated taxonomy contains twelve types of tasks split into three groups. *Data and view specification* tasks (visualize, filter, sort, derive) for exploring large datasets tend to focus on the selection of visual encodings rather than the actual visualization. *View manipulation* tasks (select, navigate, coordinate, organize) are used for highlighting and coordinating interesting items and represent the core tasks described 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* tasks (record, annotate, share, guide).

An extensive treatment of the various reusable quantitative visualizations can be found in [54]. The book presents a grammar of graphics that allows building any 2D scientific visualization from a set of simple primitives such as points, lines, scales or shapes. Recent visualization libraries built on top of D3 such as ggD3[6] or Vega[7] are following this philosophy. The next step in the evolution of LD visualization systems is to design pipelines and systems capable of exploiting the dataset structure and the underlining data structures. The next section is focused only on those systems able to visualize QB datasets.

3.2. Statistical Linked Data Visualization

In statistical LD visualization, instances will typically belong to or be associated with QB datasets. If the SLDs have a more complex structure, the corresponding ontologies or DSDs might need to be visualized as well. There are three types of *Statistical LD Visualizations systems*:

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

Basic Visualizations and Aggregations. The LOD2 project developed a *Statistical Workbench* that reflects various phases of the statistical LOD consumption cycle, e.g. triplification via CSV2DataCube, validation through the RDF Data Cube Validation tool and visualization with CubeViz[15,44]. *CubeViz* [44] 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 tools to manage the statistical LOD lifecycle [26] and includes components for Extract-Transform-Load (ETL) operations (Grafter framework), data conversion (TARQL adaptation) and data publishing (D2RQ extensions). It also contains tools for consuming the data: the OpenCube Browser for table-based views, an R package for statistical analysis, a widget for slicing data cubes, a catalog management component and a tool for interactive map-based visualizations. *Vital* [12] uses visualizations to help in the analysis and debuggging process for QB datasets publication. The automated *Visualization Wizard* described in [36] 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 [4] and Google Charts). Another paper related to the same project [24] presents the *Linked Data Query Wizard* which uses a table-based approach to selecting query results from QB datasets, and classic chart types or mind maps to visualize the results. Ba-Lam Do [13] 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 rather than as a specification and almost always come up with some changes to it; b) the challenge of creating tools for consuming statistical LD.

Dashboards. There are several dashboards based on the RDF Data Cube format (QB) and its predecessor, the Statistical Data and Metadata eXchange (SDMX) format - the ISO standard for statistical data representation currently used by large institution such as the United Nations, Eurostat, the International Monetary Fund, and the World Bank. The dashboards of Jern [25] and Hienert [23] used SDMX as they were built before 2012. One of the first QB dashboard examples by Sabol et al. [40] extends earlier work [36,24] and allows brushing over multiple coordinated visualizations. Sabol's paper analyzes two scenarios (search and analysis over LOD, analysis of scientific publications), describes the underlying workflow, and the resulting visualizations implemented in the extensions of the *Visualization Wizard* tool. The framework presented in this article is based on a *Multiple Coordinated View* (MCV) architecture to synchronise multiple visualizations [45].[8] This approach addresses several challenges identified in the Semantic DSS study [2].

Even though not directly related to visualizations, the work of Kämpgen and Harth [28] focused on interrogating multiple QB datasets via the OLAP4LD framework, which can be used as a starting point for delivering data to complex dashboards.

[6] benjh33.github.io/ggd3

[7] trifacta.github.io/vega

[8] The *Media Watch on Climate Change* is a news and social media aggregator on climate change and related environmental issues, which serves as a public showcase of the presented MCV approach (www.ecoresearch.net/climate).

Linked Data Platforms (LDPs).[9] The idea behind LDPs is the delivery of data in various formats using REST APIs (Application Programming Interfaces following the Representational State Transfer standard). Some platforms also allow to build full-featured interfaces that can contain maps or pictures, typically using templating solutions such as Velocity,[10] Elda,[11] Carbon LDP,[12] Apache Marmotta,[13] Graphity,[14], LDP4j [19] and Virtuoso[15] are several exponents of this trend. Many of the applications developed following LDP best practices[16] include maps or other types of visualizations. The main reason to include them as a separate type of visualization is the fact that many of these platforms are used to publish QB datasets.

In conclusion, many visualization workflows are geared towards creating simple charts, and little effort (with the exception of MCV dashboards) is dedicated to complex analytic solutions. Without combining datasets and synchronizing multiple visualizations with the integrated repository, it will remain a challenge to clearly present complex use cases like those described in Kämpgen and Harth's work or at the beginning of this article. The next section presents a set of principles and a workflow to support the creation of complex visualisations from statistical linked data.

4. Visualizing Statistical Linked Data

Before describing the design of visualizations following QB principles, this section outlines the workflow of statistical LD visualizations, as well as the tasks and visual metaphors involved.

4.1. RDF Data Cube Visualization Principles

The principles outlined in this section do not fundamentally change those presented by Dadzie and Rowe [11], but extend them in the context of visualizing statistical LD. The goal is to create a set of linked views for analyzing the relations between slices from multiple datasets in order to identify correlations and other patterns. The iterative process of defining these principles started with the design guidelines for QB datasets, iteratively expanding and refining them during dashboard development. The following list summarizes these principles for visualizing statistical LD.

- **Linked Views for Statistical LD**. Visualizations should reflect the linked nature of the data and support switching between visualizations when navigating the underlying datasets, or at the very least reflect changes across several visualizations on a single screen using multiple coordinated views technology. We refer to this principle as **Linked Visualizations for Linked Data**, 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.
- **Integrate 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 can be used to display the results of a statistical test.
- **Visualize Slices Instead of 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 instead of datasets*. 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 bed nights 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 bed nights. 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 spe-

[9] www.w3.org/tr/ldp
[10] velocity.apache.org
[11] www.github.com/epimorphics/elda
[12] www.carbonldp.com
[13] marmotta.apache.org
[14] www.github.com/graphity
[15] virtuoso.openlinksw.com
[16] dvcs.w3.org/hg/ldpwg/raw-file/default/ldp-bp/ldp-bp.html

cific context (together with text or data, for example) also increases the value of the information presented to the user.

– **Apply 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, it is difficult to predict 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 or a single view which is updated whenever the user wants to change the query. Implementing switching mechanisms for the fixed dimensions allows for flexibility in the choice of slices to be visualized.
– **Highlight Particular Observations**. The "Highlight links" principle from Dadzie and Rowe's work[11] 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. Charts could use heatmaps to emphasize importance, while in tables row/column coloring or fonts can achieve a similar effect.
– **Normalize Indicator Values**. If one wants to plot multiple indicators in order to observe correlations between them, it is not only useful, but recommended to normalize their values so that they can all be displayed in the same plot. Normalization is often a part of visualization processes, but in the case of statistical indicators it is almost always needed. In fact without normalization SLD visualization would not be possible at all. Indicator selection is extremely important, as even though after normalization any indicator can be plotted against any indicator, perhaps it would be better to plot only things that mean something in a certain use case.
– **Provide Highly Customizable Temporal Controls**. The temporal dimension plays an important role in SLD, therefore temporal controls should provided but not at the expense of overcrowding the interface. While this might seem obvious, it is a principle that was overlooked when creating the initial SLD dashboard for the ETIHQ project (see Section 5.1), since time was being perceived important just for slicing datasets. However, when one needs to understand data properly, time is essential, as things like seasonality, peaks or valleys can only be understood when using different time perspectives.
– **Extract and Share**. One of the main principles behind LD is its accessibility in multiple machine-readable formats. A quick way to achieve this through visualizations is to export the data slices into various formats. Customizable image export functions should support the dissemination of new research insights, for example, and reflect the last principle we propose, that of *extracting and sharing visual knowledge*.

4.2. Architecture and Workflow

Many state-of-the-art applications emphasize automation and reuse - creating, using and replacing visualizations as part of an iterative process. Such a lifecycle can be expressed as a series of visualization pipelines [13,26], which requires developers to follow certain workflows. When visualizing statistical LD, such workflows will necessarily include both LD tasks (selection of indicators, ontology alignment, etc.) and visualization tasks (data wrangling, interaction, etc.).

Building on best-practice examples reported in the literature, we propose a workflow for creating visualisations that follows the logical sequence of developing statistical LD applications. We have closely followed these steps when implementing the dashboards described in Sections 6 and 8.

1. **Requirements** need to be well-understood to produce a good narrative and a first set of visualization ideas (even if there is limited information about how the data looks like at this point). It is recommended to describe these requirements through scenarios or user stories. A *scenario* offers a high-level view of important application features - including the motivation and research questions, important queries to be explored, example data sources, and the type of vi-

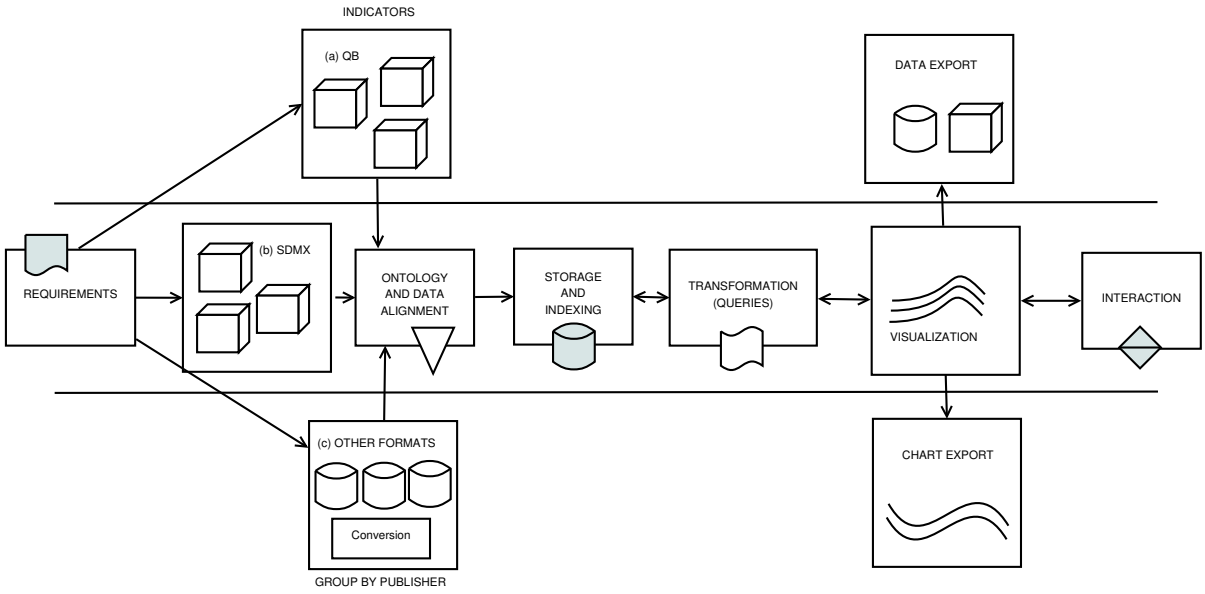

Fig. 1. Generic, reusable workflow for visualizing Statistical Linked Data (QB, SDMX and other formats)

sualizations appropriate to address the research questions. Scenarios are a good fit for describing larger applications with multiple types of interaction, many data sources, and most likely significant customization efforts. *User stories* are a good fit for smaller applications and typically describe only one feature. Ideally, scenarios should be split into several user stories. During requirements collection special emphasis should be placed on how to link visualizations in the context of an application. Another important point is to understand what kind of workflow best describes a new use case, especially the changes in existing workflows required to add the new features requested by a project or client.

2. **Discovery and Indicator Selection.** If data sources were already identified in the requirements phase, this step will require to select the needed indicators and convert them to a format that is easy to use for visualizations (e.g., a flavor of RDF or JSON). The discovery of the right indicators is not necessarily a straight-forward process, as one will have to not only identify the right indicator, but also the version that is best for the application (e.g., there can be hundreds of variants of GDP indicators published by large statistical data publishers). To get the right version of the indicator one will have to check for the name, granularity (yearly, monthly, daily), geolocation (it frequently happens that some indicators are not available for all locations or that the label of the location is different from one dataset to another) or additional clues (Which GDP indicator is needed - GDP, GDP growth or GDP per capita?). In general this step almost always requires either the use of LD and domain experts (a strategy we have deployed in the initial phases of building the ETIHQ dashboard (see Section 5.1) when the experts provided a set of URIs to index), or the creation of an automated tool to discover and import the data (a strategy we have started to deploy later in the process, once we had a better understanding of the data). Grouping the indicators according to their provenance and category helped in the later stages of the process. In this stage lies the key to developing multiple types of workflows based on the data provenance and format, as it can easily be seen in Figure 1. The current article is mostly focused on the QB workflow. RDF workflows need an extra cubification step, while other formats such as *Comma-Separated Values* (CSV) require both cubification and conversion steps. While we do not explicitly show additional cubifications in Figure 1, the next steps of the workflow assume the data is in a QB or QB inspired format (e.g., a JSON representation of QB datasets).
3. **Ontology and Data Alignment.** Running a sequence of SPARQL queries can yield an abundance of data, but to create real value, the var-

ious dimensions of the datasets under consideration need to be analyzed, aligned, augmented or aggregated in order to fit particular visualization scenarios. The ontology alignment phase only refers to the analysis and alignment of the data. There are a number of important steps that need to be taken into account when performing ontology alignment between QB datasets: fixing or avoiding broken or missing DSDs, failure of SPARQL endpoints, broken dumps, missing code lists. All these have to be included into alignment queries or scripts. The strategy we used in order to ease the ontology alignment process was to examine sets of RDF dumps from large statistical data publishers (e.g., examine several random World Bank or Eurostat QB datasets) and determine upfront if we needed additional data (e.g., code lists). By focusing on the publishers instead of particular datasets, we are able to easily ingest all the datasets from each publisher since they will all follow the same rules (e.g., the DSDs will follow the same format, the code lists will be similar, etc).

4. **Indicator Storage and Retrieval.** Storage addresses the problem of failing SPARQL endpoints, and coupled with effective indexing strategies in conjunction with established platforms such as Elasticsearch, Lucence or Sindice is essential when building IR applications. For publishers with large number of datasets we index the RDF dumps. We prefer to index triples due to the advantages offered by modern search engines like Elasticsearch (speed, availability, simple document structure). Familiarity plays an important role in this phase, as it is important to choose triple stores and search engines that are known by the development team.
5. **Transformation.** This is one of the most important steps of the workflow, as it allows to specify data wranglers [29] (scripts that transform data into formats suited for particular visualizations), queries or aggregations. Since the data items were already indexed using a search server there was no need for data wrangling scripts as the indexer already performs this mapping function. However, in this step we wrote the queries and aggregations. We view the transformation step 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 [21].
6. **Visualization.** Once the data is indexed, any query has to lead to at least one visualization. As opposed to approaches that focus on creating a single chart [40], the goal was to generate a set of linked visualizations. This simplifies the process for first-time users, who do not need to choose a particular representation of the data representation, but can look at the data from different perspectives. Since data has already been aligned in a previous step, all the visualizations got similar input regardless of provenance. Each visualization module contains all the functionality one would expect from a visualization grammar, therefore we view them as the second part of Heer and Shneiderman's data and view specification (visualize, sort) [21] in our implementation. While it might not seem important at this stage, the taxonomy used in the interfaces (e.g., entries in the menus) needs to be clear enough for the users so that they do not need to experiment too much, otherwise they might perceive the learning process of the tool as a complex task.
7. **Interaction.** An interaction layer (selections, zoom, pan, transitions, synchronization) is usually built on top of the visualization layer. Some interactive features are already built into most of the visualizations (e.g., selections, tooltips), but the features found in this layer are those that are essential to the global look and feel of the interface. This level corresponds to Heer and Shneiderman's view manipulation [21].
8. **Reuse and Sharing**. These processes can occur on multiple levels, from the indicators or indexes, to specific charts, APIs or entire platforms. Reuse should be an integral part of the design process, parts of it corresponding to process and provenance in Heer and Shneiderman's taxonomy [21]. Users should be able to share both the visual results, as well as the underlying datasets. *Chart Export* (PNG, SVG) and *Data Export* (CSV, XLS) create the possibility to easily automate reporting (a feature that is essential for business analytics), while also offering users the opportunity to quickly share their findings.

Due to the increased specialization of certain layers (e.g., alignment or interaction) and the wide variety of choices when it comes to storage and indexing, by using the previous steps as a guideline, one can implement a variety of workflows.

The next section will introduce several use cases 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 previously mentioned principles and steps.

5. Decision Support Use Cases

The main strength of LD technology lies in the simplified integration of various data sources either by aligning identical entities (e.g., statistical indicators, people, organizations, locations), or by explicitly stating the relation between similar things (e.g., one statistical indicator being narrower than another).

The following use cases present independent visual analytics platforms for different domains, which integrate statistical linked data with real-time content streams from news and social media channels.

5.1. Tourism Domain

Tourism analytics is a complex field drawing on different statistical data sources (Eurostat, World Bank, etc.) and a wide range of indicators incorporated from these sources (bed nights, arrivals, capacities, etc.). The ETIHQ project [42] investigated such sources and their value for visual tools and real-world decision support scenarios. Typical users in such scenarios are tourism professional (e.g., DMO managers, travel consultants, researchers) interested in questions related to seasonality, country and city profiles during peak tourist season, points of interests, arrivals in tourism destinations, or number of occupied bed nights. Tourism professionals will be interested in nuanced answers to such questions in order to better understand why tourists choose a certain destination in a given time period.

Many existing tools do not support the creation of scenarios for visualizing linked models as part of a unified view, or to easily reuse visualizations by changing their input data. Answering complex questions, however, often requires combining heterogeneous indicators from multiple sources. One has to specify not only the possible combinations of dimensions and measures within a visualization, but also the temporal granularity of the datasets (e.g. monthly, weekly or daily data points), their provenance, or statistical tests that are needed to validate the underlying models. To combine this heterogeneous data into meaningful visualizations, visual tools need to use MCV or similar design patterns to synchronize multiple visualizations [45].

To specify requirements in line with the scope of the ETIHQ project, we have initially conducted structured interviews with colleagues from the *Tourism and Service Management* and *Applied Statistics and Economics* departments from MODUL University Vienna, and also conducted a practitioner's survey [43]. The evaluation of the interface described in Section 7 took place after the project had ended - its results were instrumental for designing the telecommunications dashboard prototype, as outlined in the next section.

5.2. Telecommunications Domain

The effective integration of structured and unstructured data from multiple sources, both open and proprietary, is of particular importance in the telecommunications industry. Pursuing such an integrated approach, the ASAP Project (see Section 8) collects and annotates the public dialog about regional and national events in the form of Web documents and social media content, and combines the resulting repository with *Call Data Records* (CDRs) related to voice, SMS and mobile traffic, aggregated and fully anonymized to preserve customer privacy in line with European privacy protection laws.[17]

A telecommunications analyst who wants to compare CDR data across cities, for example, can use the aggregated representations of online media media coverage from the observed regions, and correlate peaks in the number of calls with co-occurring events such as music concerts, sports events and political campaigns. Statistical indicators from the respective cities can help analysts understand related geopolitical trends such as migration, an aging population, or a decreasing Gross Domestic Product (GDP).

To cope with the requirements of such scenarios, a state-of-the-art visualization engine and dashboard needs to include not just a set of appropriate visual methods, but also components that support: (i) the parallel processing of a wide variety of data types, including semantic data types like geographic location, sentiment, timestamp, etc; (ii) the remix of data from a wide variety of data sources regardless of domain, type (structured or unstructured), or provenance; (iii) and the possibility to extract aggregated statistics of the most important entities, and means to select, sort and summarize the data accordingly.

[17] ec.europa.eu/justice/data-protection/article-29/index_en.htm

5.3. Classification of Decision Support Scenarios

To better structure use case descriptions, 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 telling different stories, and mix the visualizations according to the hypothesis we want to check, but also with respect to data provenance.

The **1:1 Scenario** *(one indicator, one source)* is the most common case that inspects one indicator from a single source, e.g. showing the TourMIS[18] bed nights indicator over a period of time. Showing 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 (e.g., United Kingdom and 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.

The **N:1 Scenario** *(two or more indicators, same source)* allows inspecting multiple indicators from the same source - e.g., by displaying bed nights and arrivals from the same market to a destination one could infer the percentage of the arriving tourists who slept in hotels. It is rarely used in practice, but it is useful when we need a list of all indicators related to a certain topic from a single source (e.g., we want to know which type of arrival indicators appear in TourMIS - arrivals inside the city, arrivals at city borders, arrivals at hotels, etc) or when we need correlations between indicators from the same source.

The **1:N Scenario** *(one indicator, multiple sources* can send the wrong message to the user, but can be of interest for dataset publishers. 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). 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 (or comparable) semantics for its dimensions. This scenario could sometimes 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 there could be differences between the data collection methodologies.

The **M:N Scenario** *(multiple indicators, multiple sources)* covers the most interesting cases, often addressing interdisciplinary questions such as: How are the arrivals from a certain market influenced by the GDP growth in a market country? Do CO_2 emissions of a destination city affect its arrivals per capita? Varying the settings of an indicator can reveal interesting correlations - e.g., comparing performance on city vs. country levels, or investigating seasonal variations.

From a tourism research perspective, cross-domain indicator comparisons are the most relevant cases. LD technologies support integrated visualizations that are difficult to obtain by means of traditional database systems. 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, 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. Visual Analytics Dashboard

The visual dashboard[19] (Figure 2) is a visual semantic DSS that uses multi-domain knowledge in tourism. The dashboard combines information from TourMIS, World Bank and EuroStat. Its design is based on the scenarios discussed in Section 5. It currently allows decision makers to select and concurrently visualize tourism, economic and sustainability indicators, though the number of indicators can be extended to any number of domains of interest for which statistical LD exists. 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 (1:N Scenarios) are not present in this dashboard.

[18] www.tourmis.info

[19] etihq.weblyzard.com

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	**1:1 Scenario**: 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?*	**N:1 Scenario**: 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 bed nights)*
2 (+) sources	**1:N Scenario**: 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?*	**M:N Scenario**: 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/bed nights in one (or more) cities (e.g., Vienna vs. Amsterdam)*? – *e.g., How does tourism impact the environment of the host country?*

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.

After discussing the design of the scenarios that were important for this dashboard, we will examine how each component implements the workflow from Section 4.2.

6.1. The Linked Data Layers

In order to implement the use cases described in Section 5, one needs access to several indicators from various data publishers. The tourism data we have used represents the dumps of an updated version of TourMISLOD [41] named ETIHQ which contains tourism data about arrivals, capacities, bed nights, 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[20]. Some details about the publishing process of these RDF dumps can be found in [8,9].

Currently several issues need to be solved by anyone trying to build large scale RDF or Linked Data visualization engines: (i) SPARQL repositories still have serious availability and scalability issues and in order to federate data one will have to replicate all the needed datasets, vocabularies and Knowledge Bases locally; (ii) somewhat related to the previous issue, in order to be able to run queries against data from multiple sources one either has to perform data matching on the fly or convert those sources to a common format; (iii) the data matching is complicated by the fact that dataset designers do not strictly follow the rules from guidelines, therefore in the case of RDF Data Cubes, we frequently have missing code lists / dictionaries or DSDs, object properties that are labeled heterogeneously [56]; (iv) visualizations are usually realised with JavaScript libraries like D3, which require JSON-based formats in order to quickly process and visualize any type of data. These issues suggested indexing the data rather than to provide the users with on-the-fly integration and visualization of LD sources, as all operations should take less than a second if they are to be integrated into the portal.

As already noted, currently the QB standard is taken mostly as a recommendation, some implementations skipping code lists or data structure documents (especially if they only need to publish one dataset) or using their own heterogeneous naming conventions for object properties. As explained in [56], if location is named through geo, location, geolocation and many other names in several datasets, a matching system will have to take into account all these variations. Standardizing naming conventions for RDF Data Cubes would be one way to address some of these issues, but it would only work if the guidelines would be strictly enforced (e.g., via custom validation). Perhaps even more troublesome is the fact that some elements of the QB vocabulary like slices or observation groups are rarely

[20] www.270a.info

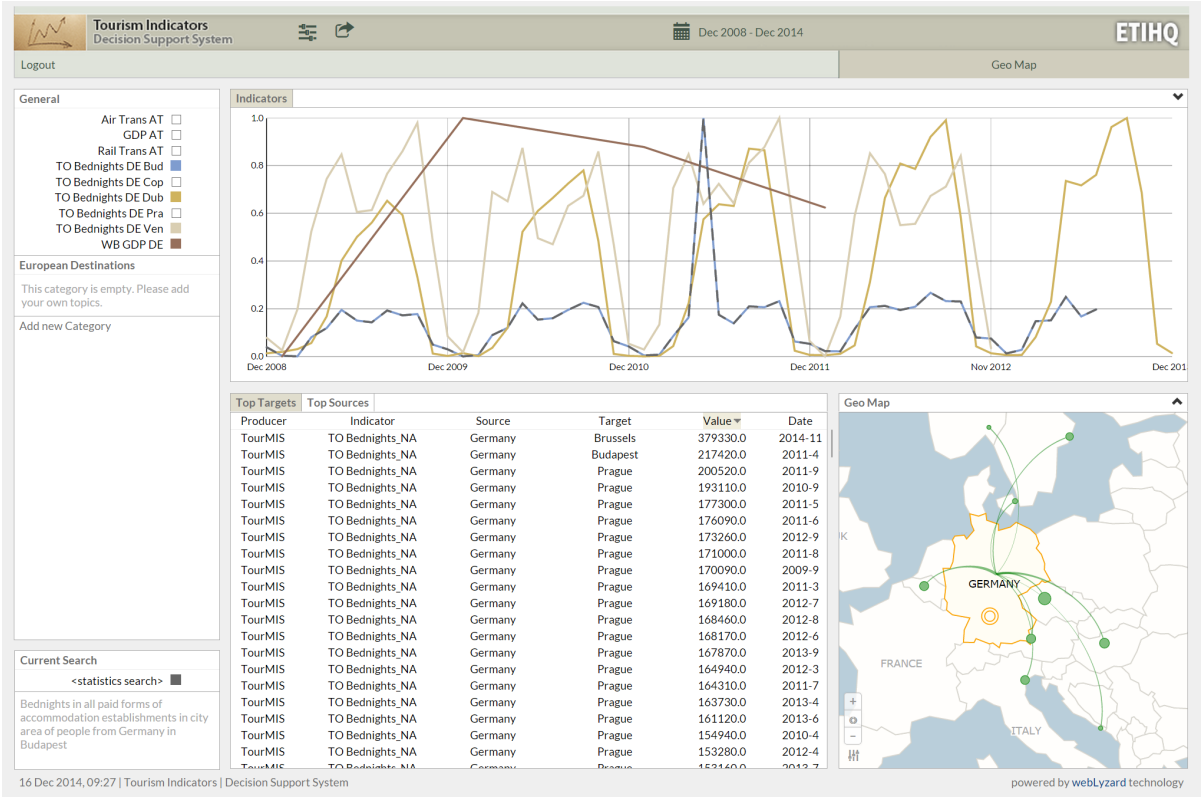

Fig. 2. The ETIHQ Dashboard showing bed nights occupied by German tourists in various European destinations (Budapest, Dublin, Venice), plotted against the GDP growth of Germany

used. While this perhaps makes sense for slices which can be created automatically later, behind the existence of an observation group there might be reasons that are not immediately obvious if not documented. For example, medical data from certain months or years can be published as an observation group because there was an epidemic in a set of countries. In such a scenario we face a loss of information if the observation group is missing.

While RDF Data Cubes do offer a simple way to merge lots of datasets together and create large data cubes with statistical indicators, this is true only if the data publishers follow the specifications point by point. The fact that some components such as code lists are optional and not necessarily well-understood leads to additional complexity. We have for example encountered several situations related to code lists: (i) they do not exist, which is not a problem since they are optional; (ii) they exist, work fine and respect the specifications; (iii) they exist, but they contain ambiguous names or URIs (which should not happen) - therefore requiring a pre-processing step since it cannot be anticipated what they will contain. In fact, some small changes to the specification of the optional components of the RDF Data Cubes and a better validation process for these components are needed before on-the-fly validation and visualization in a reasonable amount of time (sub second) can be achieved, especially if a system needs to be able to integrate any kind of SLD source. We insist on the sub second loading times to avoid suboptimal user experiences.

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 return all the URIs that respect a certain pattern, for example) and run them against the dumps collected from the three services. As a final solution, we indexed the data using a search server (Elasticsearch) and creating an LD indexer API that gets the data from all the data sources. The indexing service we created provides all the functionality for the LD layers we envisioned (Selection of Indicators; Ontol-

ogy and Data Alignment; Storage and Indexing). As long as the running time of SPARQL queries is several seconds, we will prefer the fastest method of indexing data and recomputing the queries each time instead of classic data cube methods like full or partial materialization of cuboids (slices in the current QB terminology). In fact even though full or partial materialization of cuboids is certainly useful, it is not always needed. A common use case where it is not really needed is creating a new index on top of the current indexes. Creating derived indicators, a central problem in Statistical Linked Data, requires complex computations, therefore in general it is better to use the full power of a programming language like Java and some of its DSLs instead of plain SPARQL. Since we do plan to include derived indicators in a future release, this was an additional reason to consider indexing.

After indexing the datasets, each observation corresponds to one document. The document structure for a QB observation only takes into account the essential information that needs to exist in a dataset so that it can be visualized: the observation value, the unit of measure, the geographic location (if it exists), and so on. This allows indexing a huge number of datasets from many publishers therefore enabling the creation of scalable visualization solutions.

Since the URIs from Eurostat and World Bank published in the 270 Linked Data Space [21] are well-designed, for the *discovery and selection of indicators* all that is needed to find an indicator is to have an idea about the name or part of the name of the desired indicator. If the indicator name and URIs are known, then it is sufficient to directly provide the URIs for the new datasets. In the first phase, the indexer will harvest all triples from that location that match the selected 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 have not offered enough information for a visualization. Some additional tasks that are performed are usually those related to *ontology alignment*. One such example of alignment is the geospatial alignment performed by the indexer: Geonames [55] and DBpedia [32] URIs are used instead of the names of the actual locations, as the real names of the location might suffer from various issues such as 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 kept all the information (including the links) from the actual RDF dumps so that any observation or slice can be recreated if needed. From the first set of indexed datasets we have extracted a QB-inspired JSON data format in order to ease the validation of further datasets. The required fields of this format are those expected to be found in any QB dataset (dataset, observation URI, observation value, date, etc.), while optional fields can accommodate dataset-specific information such as geographic location or the unit of measurement

The added information such as granularity or added 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 some of the functionality typically found on a transformation layer.

The first version of the indexer contained small functions that allowed indexing any type of dataset or code lists from a certain publisher (e.g., Eurostat, World Bank, TourMIS). This was possible due to the fact that each publisher follows the same style of dataset design for their datasets (e.g., if they do not use code lists, none of the datasets from that publisher will have references to code lists). Therefore by writing several lines of code we were able to automatically index hundreds of datasets from a single publisher. While this worked well, we still needed new data formats from time to time (date or time formats that were not included in the initial list, for example), as almost each dataset producer only loosely followed the W3C Recommendation when designing RDF Data Cubes and introduced small variations. Due to this fact, a custom API has been developed to provide third-party access (see Section 8).

The functionality for all these layers (*selection of indicators, ontology alignment*) is included into a single software package that was initially written in Python and later rewritten in Java for better performance.

[21] www.270a.info

6.2. 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).

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

The *reusable visualizations* are written in JavaScript with jQuery and d3.js, following the conventions of the d3 reusable chart pattern[22]. All visualizations are presented in a single-screen interface, synchronized using the multiple coordinated views design pattern.

The two layers that host the interface components are the *visualization* and *interaction* layers. In reality they often cannot be separated, as often certain types of interaction are easier to implement directly into a specific visualization, as opposed to external modules. It also helps to present the workflow that needs to be followed when constructing particular dashboard visualizations.

The current dashboard is targeting analysts and decision makers. This can be inferred directly from the use cases presented in the previous section. A Destination Management Office (DMO) 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. Some of the functionalities of the dashboard were created with researchers in mind (e.g., data export).

Adding and Visualizing Slices. The user can start exploring new questions by adding several slices. In order to add a slice one will have to choose a date interval (via the calendar button from Figure 2), then proceed to complete all the needed information about provenance and dimensions in the Advanced Search dialog and add it to the General menu.

A good method to start could be *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 gear icons button in the *General* pane (Figure 2) will uncover the 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 2, because the display space for menus will always be limited. Generally, we recommend that the names consist of the abbreviation of the indicator's data source (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 the shown example, these would be DE for Germany and Bud for Budapest). So, for this example indicator we provide the *TO Bednights DE Pra* name. While some users might not adopt this convention (indeed some of the users who participated in the evaluation described in the next section have not), it is nevertheless good to provide guidelines about this naming convention in the tutorials or user manuals.

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 define the topic. By hovering over the new topic and clicking the gear icon on the right, the chart view in the top-middle pane of the interface will be replaced with a dialog field that allows defining the topic, as shown in Figure 3. 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, *Save* will close the dialog box.

The *General* pane, the *Advanced Search* dialog, and the date selection mechanisms, allow the users to create most of the operations from the *data and view specification* layers suggested by Heer and Shneiderman [21]. The *General* pane allows to *filter* the indicators and *sort* them via the advanced search 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

[22] bost.ocks.org/mike/chart

Fig. 3. Advanced search dialog for creating slices, accessible via the topic's gear icon

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 the presented 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 2). The *Targets* table, keeps the source value fixed (Germany) 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 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 areas, and therefore have a single dimension, that of the country/city of interest. In this case (as shown in the left side of Figure 4) 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 (arrows) is created (bigger arrows mean more tourists in the selected interval). The map from Figure 2, shows various destinations that were top choices for German tourists. For the Eurostat data (Air Transport indicator), the right side of Figure 4 (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.

Interaction. Since from the previous analysis Budapest does not necessarily stand out as a popular tourist destination for Germans (which is normal given the fact that it is not compared with anything), new topics can be added that contain Bednights of German tourists to other locations (Dublin, Venice, etc). These new topics 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 behavior 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 2). Figure 2 superimposes 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 values displayed in the chart are normalized so that it is easy to compare them.

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 such as 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

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

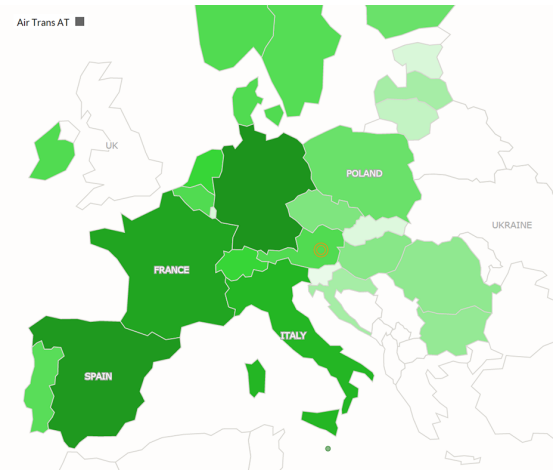

Fig. 4. Tabular and geographic map views of Eurostat data

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 [21]. We can *select* items from the tables and trigger new searches using them as parameters, or *select* various observations from the line chart and display additional information in tooltips. The geographic map allows users to see summaries of 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.

Sharing. Pushing the *Export* button, opens a side menu that allows selecting from two groups of options: *Chart Data* (XLS or CSV formats) and *Diagrams* (Line Chart, Geographic Map). They allow users 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 [21] *process and provenance* functionality, and are one of the most popular features.

6.3. Scalability

It generally takes less than a minute to load two to five typical datasets via the API, and this results in an index with the size of around a hundred MB. An index of 1000 random datasets with data from Eurostat and the World Bank has 27,2 GB and 242 million documents (= data points). We do not index data about composite geographical entities, regardless of them being real (like European Union) or made up for that specific indicator (like DE+FR or EU-25 or Vienna+Salzburg), only data that contains real geographical coordinates or corresponds to actual geopolitical entities (countries, regions, cities) or points of interests (e.g., historical sites, parks, museums). Covering the full datasets would likely result in sizes that are up to 1.5 times bigger. Elasticsearch has no latency issues even when hosting indexes several times this size. The initial load of the portal takes about 1.5 to 5 seconds. Each subsequent query and the visualization of results is typically performed in less than a second, regardless of the index size. Future versions will also support the display of composite geographical entities.

Some datasets contain millions of points, as already mentioned, but those millions of points contain all possible slices. When visualizing datasets with the ETIHQ dashboard, we already visualize slices instead of entire datasets (as explained in Section 6.2), therefore restricting the data size to only several hundreds points at most. This is the main reason why the *Advanced Search* dialog is necessary before being able to visualize new slices. A slice for the GDP of Austria for the last 10 years only has 10 points at most (if all the data points are available). Even when visualizing several slices there will be less than 120 points. If some of these datasets do have monthly data (e.g., TourMIS) there will be at most 120 points for the last ten years for a single slice and less than a 1200 when visualiz-

ing 10 such slices (the maximum number of slices that can be visualized with this tool). However, for sources with monthly or daily data, the more the time interval is increased, the chances to end up with a visualization where data is already aggregated (daily data displayed as monthly totals, and monthly data displayed as yearly totals) also grow.

7. Dashboard Evaluation

The tourism dashboard prototype showcases cross-domain data analytics functions that are feasible over datasets integrated through Linked Data. An exploratory evaluation of this prototype has been conducted and reported in [43] with the focus of understanding the usefulness of the tool for tourism practitioners. In this section we provide a summary of that evaluation and refer the interested readers to details in [43].

7.1. Evaluation Design

Evaluation goals were (a) identifying potential new functionalities that the tool could enable; (b) assessing the current usability of the prototype and deriving ideas for future usability-level improvements; and (c) obtaining indicative performance values when using this prototype as opposed to current practices for solving cross-domain data analytics.

Participants. Participants to the evaluation were selected from the two major stakeholder groups that could benefit from a cross-domain data analytics infrastructure: researchers in the tourism domain as well as tourism practitioners, working primarily for Destination Management Offices (DMO's). The 16 participants were divided randomly in two groups (Group_A and Group_B), both containing an equal number and mixture of researchers and practitioners, i.e., five practitioners and three researchers per group.

Evaluation Setup. Prior to the evaluation itself, each participant received a tutorial that explained the features of the tool, and included practical exercises to ensure a basic familiarity with the tool (e.g., how to create and define indicators, how to visualize and compare their values). Evaluations were performed at the desk of each participant and at a time that best fitted the participant's schedule. This allowed to maintain a realistic work environment and to avoid bias potentially introduced by requesting the use of a new work environment, such as in lab-based evaluation settings. Additionally, such a setup was the only option that allowed involving DMO employees from across Europe (this design setup was suitable for an exploratory study, but future baseline comparisons will consider a more controlled lab-based setup). The evaluation included the four following activities:

- **Activity 1**. Participants performed three tasks using the Dashboard and recorded the time taken to perform each task and the results they have reached. See Table 2 for an overview of the tasks as well as their assignment to the two groups. Group_A performed tasks T1-T3 with the Dashboard, while Group_B focused on tasks T4-T6.
- **Activity 2**. To gather insights about potential new uses of the tool, participants were asked to create and perform two tasks of their own using the Dashboard. They noted the tasks they preformed and the insights they gained.
- **Activity 3**. To collect information about how the evaluated tasks would be typically performed in state of the art settings, participants performed three tasks without using the Dashboard. Participants were allowed to adopt their usual data collection and analytics approach. Participants noted their findings, the time taken to perform each task as well as the approach and tools they made use of. In this activity, the role of the groups was inverted, with Group_A performing tasks T4-T6 without the Dashboard, while Group_B focused on tasks T1-T3.
- **Activity 4**. To get an insight into the usability of the tool, participants answered the ten questions that make up the System Usability Scale (SUS), the most used questionnaire for measuring perceptions of system usability [5]. Additionally, they provided feedback about the most and least useful features of the tool, as well as their recommendations for future extensions of the tool.

Evaluation Tasks. Six tasks were proposed to the participants (Table 2). To measure improvements in terms of time savings and quality of the answers, each task was performed either with the dashboard, or using traditional desktop spreadsheet applications. Averages for the time spent on each task as well as the accuracy of the provided answers were measured and compared. Participants were instructed to abandon a task if they failed to complete it within 15 minutes. This facilitated time management and followed similar rules reported in the literature [37], considering that in practice longer tasks would often be abandoned.

Table 2
Assignment of tasks to the two evaluator groups (GrA, GrB);
Y = dashboard use; N = other means, e.g. spreadsheet applications

Task	Description	GrA	GrB
1	In which month was the number of bed nights of tourists from the USA to Germany the highest?	Y	N
2	Explore possible similarities between the American economy (as indicated by GDP Growth) and the bed nights spent by American tourists in Germany.	Y	N
3	Continuing your exploration from task 2, compare how Germany and Austria have been affected (in terms of bed nights of American tourists) by the economic situation in the USA.	Y	N
4	In which month was the arrival of Japanese tourists to Vienna the lowest?	N	Y
5	Could the GDP of Japan have had any influence on the number of arrivals of Japanese tourists in Vienna?	N	Y
6	Continuing your exploration from task 5, compare how Budapest and Vienna have been affected (in terms of arrivals of Japanese tourists) by the economic situation in Japan.	N	Y

The tasks were formulated to cover two exploration scenarios. Task 1 to 3 investigate the influence of the American economy on bed nights of American tourists to different European countries. Similarly, tasks 4-6 cover an exploration of how Japanese economic indicators might influence arrivals of Japanese tourists to European cities. Both scenarios were investigated in the same time period (January 2005 - December 2014).

7.2. Evaluation Results

Evaluation results have shown that the ETIHQ Dashboard provides considerable improvements over manual approaches to answering a range of complex questions. By comparing the outcomes of Activity 1 and 3, an average time improvement of 29.76% was obtained (average task execution times without the dashboard were 5.74 minutes and with the dashboard 3.93), while answer quality, in terms of precision, was clearly inferior when using manual approaches (under 71%) as opposed to using ETIHQ (over 63% and up to 100%). Based on Activity 2 it became clear that manual approaches to cross-domain analytics are currently the norm and participants provided ideas for new tasks to be supported with the Dashboard. Details on all these results are available in [43].

A third important aspect of the evaluation was the focus on the usability of the tool and the collection of future features, based on the results from Activity 4. We report these findings and extend them beyond [43].

Based on the responses to the SUS questionnaire (first part of Activity 4), an overall SUS value of 64 was computed, which on the adjective rating scale [1] is satisfactory (OK). While this results indicate that improvements are needed in terms of system design, learnability and usability [6], it is an encouraging result for a prototype, if we also factor in that the free text comments (see below) were really positive, regardless of the participants being from the academy or industry. The fact that almost all of the participants chose visualizations and data integration as the top features of the ETIHQ dashboard (see Table 3) is a good indicator that they understood the purpose of dashboard and consider it useful.

The second part of Activity 4 was dedicated to current and future features that users consider useful, collecting the users' answers as free form text.

Most Useful Tool Features. As expected, the ability to *easily visualize slices from multiple data sources in the same chart* was considered the most useful feature (see Table 3). Normalization was also appreciated as a good idea, since all users wanted to be able to easily observe correlations. The ability to *preview* the variables or slices before saving them was considered the best feature for exploratory research. The fact that the system automatically creates multiple linked graphics like the best statistical tools currently available was also noticed by most of the users. Overall, the users appreciated the ease of use of the system and the advanced customization features. The system's export functionality was considered a nice bonus, and some users (especially researchers) use it as a first step in their data collection strategy for future research articles or projects. We have not however received such an article until the date of the current submission. The majority of users appreciated the Advanced Search and menu-based selection mechanisms.

Open Challenges. Search for indicators was considered problematic - while the ability to create customized slices is much appreciated (see Table 3), users would like to perform indicator searches directly, without the Advanced Search dialog. Tables were mentioned as well, but contrary to what some of the survey

Table 3
Aggregated user feedback on specific dashboard features

Feature	Count	Positive Comments
Charts	14	Clear information visualization
Data Integration	14	Seamless integration of data sources
Comparison (Line Chart)	12	Benchmarking with several crucial indicators
Slice Creation	5	Option to create and customize variables
Automatic Plots	4	Automated rendering and scaling of the plots
Feature	**Count**	**Negative Comments**
Temporal Controls	4	Additional options to select time intervals
Search for Indicators	4	Inability to search without defining the source and the target
Country Totals	3	The computation of country totals should be managed automatically
Missing Data	3	Important countries do not deliver data
Tables	3	Cannot use the table to identify the most relevant data

participants have indicated, they do allow users to sort the data and identify the most important elements. Perhaps without visual highlighting via colored columns or bold typeface, this is not obvious to first time users. Table 3 might lead to the assumption that data is missing and datasets were not fully indexed, but it is a problem that stems from the ingested datasets themselves. Future work could address this feedback by additional validation steps, automatically identifying missing values in other sources, or integrating LD Quality Assessment models [57].

Technology Roadmap. Users recommended to use domain-specific terminology - e.g., *market* and *destination* for common tourism indicators, instead of generic terms such as *indicator* and *source*. Some changes to the data aggregation features were suggested to simplify the workflow: aggregated annual data; fewer data items in the top targets or top sources tables; adding controls to change data granularity in the trend charts; improve time selection mechanisms.

Participants also suggested to extend the system by including other data sources: news media, stock exchanges, flight connections, GDP indicators, exchange rates, and events (the latter not being a statistical dataset). Another requested upgrade were additional analytic functions (e.g., calculation of explicit correlations; regression analysis; more data summaries) to complement the current visual comparison, and to reduce the need to export data to other tools for detailed mathematical analysis. Other users suggested minor user interface changes such as different fonts or geographic base layers, and an interactive tutorial built directly into the tool.

8. Integrated Analytics for Structured and Unstructured Data

The ASAP FP7 Research Project[23] develops a dynamic execution framework for scalable data analytics based on structured and unstructured data from a range of sources, open as well as proprietary. This includes automatically annotated news and social media content, statistical indicators from linked data sources, and anonymized *Call Data Records* (CDRs). The conceptual development of the ASAP dashboard to analyze data across these sources benefited from the evaluation results summarized in the preceding section. The dashboard is implemented as part of the telecommunications use case (Section 5.2), paying particular attention to (i) on-the-fly data composition and visualization, (ii) the display of temporal data and interactive controls to quickly select desired time intervals, and (iii) the provision of open APIs for uploading datasets and annotations, querying the knowledge repository, and embedding visualizations into third-party applications.

Temporal Controls. Addressing user feedback gathered during the evaluation, we have added several new alternatives for temporal slicing. Perhaps most important for SLD is the *Granular Overview Overlay* described below, as it allows users to understand the temporal distribution of large datasets. It is complemented by a classic timeline and an interactive date range selector, which consists of a time-span bar in conjunction with a sliding window to quickly get a visual overview and select the desired time frame.

[23] www.asap-fp7.eu

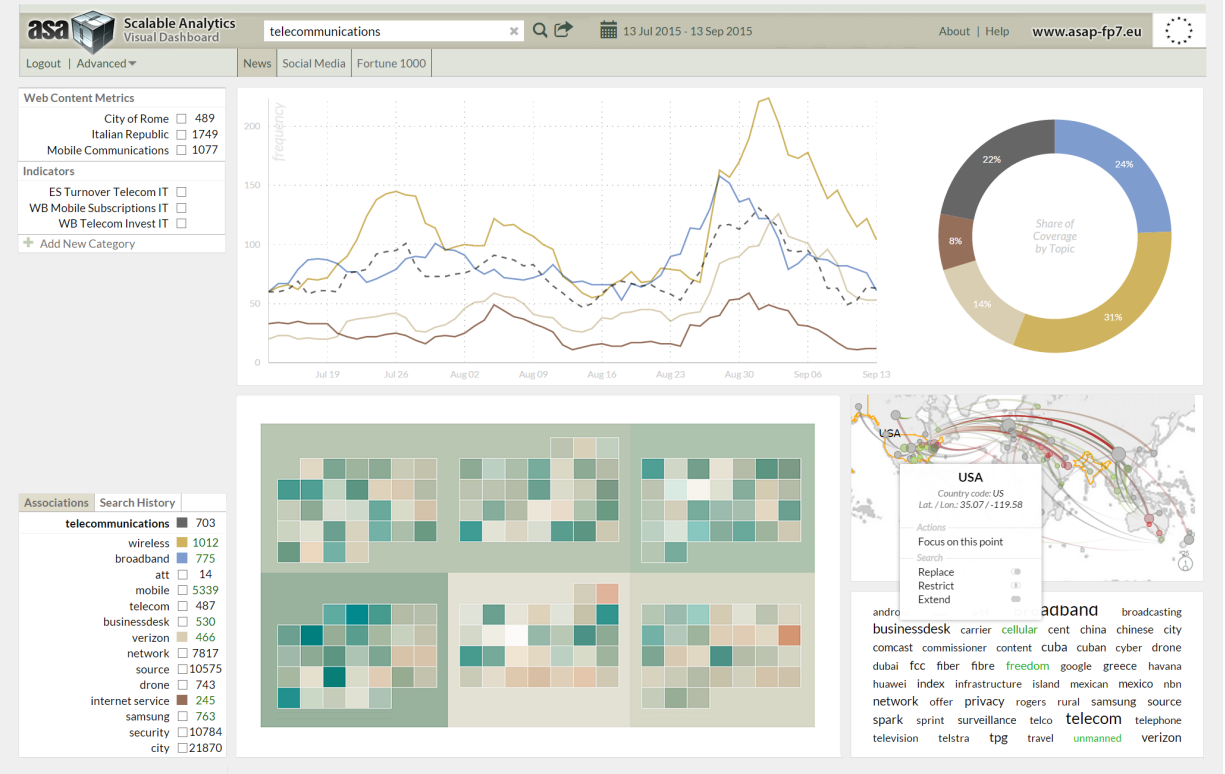

Fig. 5. Prototype of the ASAP dashboard for the integrated analysis of unstructured Web content and structured telecommunications data

Geographic Map. A revised module to render geographic projections in different resolutions benefits from the adaptive context menus shown in Figure 5. The new implementation supports the creation of customized maps and multiple base layers including a rich set of styling and formatting options for data layers, points of interests, and labels.

Adaptive Tooltips. The statistical visualization components were updated to reflect the design principles of the webLyzard Web intelligence platform[46, 47]. Adaptive tooltips and context menus enrich the functionality of the statistical components and ensure a unified user experience. Based on the current context of the analyst, for example in the form of a country shape or point of interest, the tooltip of the geographic map displays filtered information and a context menu to either drill down or extend the search. If the tooltip corresponds to a country, it will display basic information about that country (see Figure 5). Highlighting functions include visual cues to show specific groups of data points when users hover over related components in the ASAP dashboard, or rule-based color coding of these data points. The adaptive tooltips take into account the size of the container and display fewer items if a component is minimized.

Granular Overview Overlay. The temporal distribution of large datasets is shown using a new version of the GROOVE visualization introduced by Lammarsch et al. [31] and seen under the line chart in Figure 5. The current prototype comes with integrated focus and context based on granularities and recursive pattern arrangement. It is particularly useful for visualizing daily data from third parties, and for identifying peaks or changes in behavior triggered by various events. Each data element is presented by a square that might be only the size of one pixel, but automatically expands if more space is available. The pixels (or squares) are arranged according to the days of a year, similar to a calendar. The example from Figure 5 shows six calendar months (each row corresponds to a quarter), positioning the days of each month (each day with a white border) inside a larger rectangle (the larger areas without border). Color-coding schemes are used for daily occurrences, sentiment, peak values. Monthly averages are mapped to the colors of the rectangles that surround the days and represent the months.

Data and Visualization Services. The *webLyzard API Specification*[24] bundles together several interfaces to create a uniform framework for the rapid integration of multiple data sources into a scalable visualization processing pipeline following a *Visualization as a Service* (VaaS) approach. The goal is to provide a unified interface through which to expose all data and visualization services. The *Document API* ingests unstructured text data, for example crawled Web pages or digital content from document management systems. The main objects are *Documents, Sentences* and *Annotations*. This API can be used for sharing documents regardless of their provenance, as well as annotations from knowledge extraction services including sentiment analysis [51] and named entity recognition [52]. The *Statistical Data API* ingests structured data following the RDF Data Cube philosophy. This API supports the full workflow presented in Figure 1. The *Search API* returns a set of query results in the form of unstructured text documents or time series data. The *Embeddable Visualization API* provides a mechanism to integrate visualizations into third-party applications, typically based on the results of a search query.

9. Conclusion and Future Work

Visual tools for representing statistical linked data (LD) can support policy experts and decision makers in a wide range of domains such as telecommunications, travel and tourism, financial markets, health care services, or sustainable development. Facilitated by the adoption of LD technologies, applications that seamlessly integrate and visualize statistical LD from multiple sources have started to appear. The large-scale integration of statistical LD technologies at semantic and syntactic level is still at an early stage and calls for improved methods to align, link and visually explore datasets from heterogeneous sources.

This article describes innovations from two European research projects that advance the state of the art in terms of: (1) workflow and design principles to develop statistical LD visualizations for heterogeneous data from multiple sources, (2) use cases and scenarios for visualizing statistical data, (3) visual DSSs that support these scenarios across application domains, and (4) multiple coordinated view technology for LDPs that embeds data analysis and visualization processes in a flexible and reusable framework.

[24] www.weblyzard.com/api

The presented workflows and principles can also be applied to other types of data: scientific data (often statistical in nature, but not always published using QB standards), personal data (e.g., curricula vitae or FOAF profiles), or historical data (e.g. comparing trends in different historical periods). At this stage, individual research communities develop their own guidelines, which eventually should be merged into a general set of principles for visualizing any kind of LD content.

For the use cases, we acquired statistical indicators from international organizations. One of the main benefits of using such LD sources is access to a large number of indicators (the Eurostat endpoint alone provides 6538 datasets) covering a wide range of topics (telecommunications, economy, ecology, travel and tourism, health, etc.). Publishers typically provide indicators in standard formats such as *RDF Data Cube* (QB) and *Statistical Data and Metadata eXchange* (SDMX). The datasets are interlinked, even though their alignment might represent a challenge due to heterogeneous property labels or deviations from the W3C recommendations.

QB datasets and related formats support structured queries that consider complex hierarchies exposed as code lists (e.g., geographical locations), which is not possible when using open data provided in *Comma-Separated Values* (CSV) format. A side-effect of using RDF Data Cube was the creation of a QB-inspired JSON data format for including data into the webLyzard visualization engine that can be reused for any type of statistical dataset. The required fields from this format are those that are generally used to describe datasets and observations with the QB vocabulary (dataset, observation URI, observation value, date, etc.), while optional fields can accommodate dataset-specific information such as geographic location or the unit of measurement.

The validity of the underlying datasets is of particular importance and a fruitful avenue for future research. This article is based on datasets published by trusted third parties, but large-scale LD analytics solutions require verification techniques and certification procedures to address the open nature of LD and ensure data quality including assessments of veracity and conformity with security standards. *Data composition and visualization* services will enable users to create new indicators on-the-fly, compare them with similar values from other indexes [17], and integrate them into complex analytical models - to automatically verify knowledge [49], for example, or to identify rumors spread via social media [3].

Future work will add decision support metrics extracted from news and social media coverage to these analytic models, resulting in novel solutions for the integrated management and analysis of statistical LD. This will not only unlock significant commercial opportunities, but also enable us to better understand (and act upon) systemic issues on a society-wide scale, for example when addressing environmental problems or pursuing sustainable development goals [45].

Acknowledgements

The research presented in this article has received funding from the European Union's *7th Framework Programme for Research, Technology Development and Demonstration* as part of the ASAP[25] and ETIHQ[26] (PlanetData) projects under the Grant Agreements No. 619706 and 257641, respectively. The authors would like to thank K. Wöber and I. Önder for their tourism domain expertise, and R. Kamolov, W. Rafelsberger and T. Lammarsch for developing some of the presented visualizations.

References

[1] A. Bangor, P. Kortum, and J. Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *J. Usability Studies*, 4(3):114–123, May 2009.

[2] E. Blomqvist. The Use of Semantic Web Technologies for Decision Support - A Survey. *Semantic Web*, 5(3):177–201, 2014. 10.3233/SW-2012-0084.

[3] K. Bontcheva, M. Liakata, A. Scharl, and R. Procter. 1st International Workshop on Rumors and Deception in Social Media: Detection, Tracking, and Visualization (RDSM-2015). In A. Gangemi, S. Leonardi, and A. Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 945–946. ACM, 2015.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D^3 Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. 10.1109/TVCG.2011.185.

[5] J. Brooke. SUS-A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, 189(194):4–6, 1996. 10.1.1.232.5526.

[6] J. Brooke. SUS: A Retrospective. *Journal of Usability Studies*, 8(2):29–40, 2013.

[7] J. M. Brunetti, S. Auer, R. García, J. Klímek, and M. Nečaský. Formal Linked Data Visualization Model. In E. R. Weippl, M. Indrawan-Santiago, M. Steinbauer, G. Kotsis, and I. Khalil, editors, *Proceedings of the 15th International Conference on Information Integration and Web-based Applications & Services, IIWAS '13, Vienna, Austria, December 2-4, 2013*, pages 309–318. ACM, 2013. 10.1145/2539150.2539162.

[8] S. Capadisli, S. Auer, and A.-C. Ngonga Ngomo. Linked SDMX Data. *Semantic Web - Interoperability, Usability, Applicability*, 6(2):105–112, 2015. 10.3233/SW-130123.

[9] S. Capadisli, S. Auer, and R. Riedl. Linked Statistical Data Analysis. In S. Capadisli, F. Cotton, R. Cyganiak, A. Haller, A. Hamilton, and R. Troncy, editors, *Proceedings of the 1st International Workshop on Semantic Statistics co-located with 13th International Semantic Web Conference (ISWC 2013). Sydney, Australia. October 11th, 2013.*, volume 1549 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[10] E. H. Chi. A Taxonomy of Visualization Techniques Using the Data State Reference Model. In *IEEE Symposium on Information Visualization*, pages 69–75. IEEE, 2000. 10.1109/INFVIS.2000.885092.

[11] A. Dadzie and M. Rowe. Approaches to Visualising Linked Data: A Survey. *Semantic Web*, 2(2):89–124, 2011. 10.3233/SW-2011-0037.

[12] E. Daga, M. d'Aquin, A. Gangemi, and E. Motta. Early Analysis and Debugging of Linked Open Data Cubes. In S. Capadisli, F. Cotton, A. Haller, A. Hamilton, M. Scannapieco, and R. Troncy, editors, *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference (ISWC 2014). Riva del Garda - Trentino, Italy. October 19th, 2014.*, volume 1550 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[13] 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*, pages 282–290. SciTePress, 2014. 10.5220/0005110102820290.

[14] 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. 10.1007/978-3-319-13704-9.

[15] 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 Berlin Heidelberg, 2013. 10.1007/978-3-642-41360-5_19.

[16] P. Fox and J. Hendler. Changing the Equation on Scientific Data Visualization. *Science*, 331(6018):705–708, 2011. 10.1126/science.1197654.

[17] J. E. L. Gayo, H. Farhan, J. C. Fernández, and J. M. Á. Rodráguez. Representing Verifiable Statistical Index Computations as Linked Data. In S. Capadisli, F. Cotton, A. Haller, A. Hamilton, M. Scannapieco, and R. Troncy, editors, *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference (ISWC 2014). Riva del Garda - Trentino, Italy. October 19th,*

[25] www.asap-fp7.eu

[26] www.etihq.eu

2014., volume 1550 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[18] V. Geroimenko and C. Chen, editors. *Visualizing the Semantic Web. XML-Based Internet and Information Visualization*. Springer London, 2002. 10.1007/1-84628-290-X.

[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] J. Heer, M. Bostock, and V. Ogievetsky. A Tour Through the Visualization Zoo. *Commununications of the ACM*, 53(6):59–67, 2010. 10.1145/1755884.1780401.

[21] J. Heer and B. Shneiderman. Interactive Dynamics for Visual Analysis. *Communications of the ACM*, 55:45–54, 2012. 10.1145/2133416.2146416.

[22] 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, *he 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 International Publishing, 2014. 10.1007/978-3-319-11955-7_50.

[23] D. Hienert, B. Zapilko, P. Schaer, and B. Mathiak. Web-Based Multi-View Visualizations for Aggregated Statistics. *CoRR*, abs/1110.3126, 2011. arxiv.org/abs/1110.3126.

[24] P. Höfler, M. Granitzer, E. E. Veas, and C. Seifert. Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints. In C. Bizer, T. Heath, S. Auer, and T. Berners-Lee, editors, *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, volume 1184 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[25] 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. 10.1007/978-3-642-04265-2_5.

[26] 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. Capadisli, F. Cotton, A. Haller, A. Hamilton, M. Scannapieco, and R. Troncy, editors, *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference (ISWC 2014). Riva del Garda - Trentino, Italy. October 19th, 2014.*, volume 1550 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[27] 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. 10.1147/JRD.2012.2188737.

[28] 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 International Publishing, 2014. 10.1007/978-3-319-11955-7_54.

[29] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In D. S. Tan, S. Amershi, B. Begole, W. A. Kellogg, and M. Tungare, editors, *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011.*, pages 3363–3372. ACM, 2011. 10.1145/1978942.1979444.

[30] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. G. Giannopoulou. Ontology Visualization Methods - A Survey. *ACM Computing Surveys*, 39(4), 2007. 10.1145/1287620.1287621.

[31] T. Lammarsch, W. Aigner, A. Bertone, J. Gartner, E. Mayr, S. Miksch, and M. Smuc. Hierarchical temporal patterns and interactive aggregated views for pixel-based visualizations. In *13th International Conference on Information Visualisation, IEEE VIS 2009*, pages 44–50. IEEE, 2009. 10.1109/IV.2009.52.

[32] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015. 10.3233/SW-140134.

[33] 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 International Publishing, 2014. 10.1007/978-3-319-13704-9_21.

[34] E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963. 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[35] 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) co-located with the 13th International Semantic Web Conference (ISWC 2014). Riva del Garda, Italy, October 20, 2014.*, volume 1279 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[36] 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 Posters and Demonstrations Track, I-SEMANTICS 2013*, volume 1026 of *CEUR Workshop Proceedings*, pages 40–44. CEUR-WS.org, 2013.

[37] F. Osborne, E. Motta, and P. Mulholland. Exploring Scholarly Data with Rexplore. In H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, vol-

ume 8218 of *Lecture Notes in Computer Science*, pages 460–477. Springer Berlin Heidelberg, 2013. 10.1007/978-3-642-41335-3_29.

[38] H. Paulheim and F. Probst. Ontology-Enhanced User Interfaces: A Survey. *International Journal on Semantic Web and Information Systems*, 6(2):36–59, 2010. 10.4018/jswis.2010040103.

[39] J. Polowinski. Towards RVL: A Declarative Language for Visualizing RDFS/OWL Data. In D. Camacho, R. Akerkar, and M. D. Rodríguez-Moreno, editors, *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS 2013, Madrid, Spain, June 12-14, 2013.*, page 38. ACM, 2013. 10.1145/2479787.2479825.

[40] 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. Benrstein, 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 International Publishing Switzerland, 2014. 10.1007/978-3-319-11964-9_20.

[41] M. Sabou, I. Arsal, and A. Braşoveanu. TourMISLOD: A Tourism Linked Data Set. *Semantic Web - Interoperability, Usability, Applicability*, 4(3):271–276, 2013. 10.3233/SW-2012-0087.

[42] M. Sabou, A. Braşoveanu, and I. Önder. Linked Data for Cross-Domain Decision-making in Tourism. In I. Tussyadiah and A. Inversini, editors, *Information and Communication Technologies in Tourism 2015. Proceedings of the International Conference in Lugano, Switzerland, February 3 - 6, 2015*, pages 197–210. Springer, 2015. 10.1007/978-3-319-14343-9_15.

[43] M. Sabou, I. Önder, A. M. P. Braşoveanu, and A. Scharl. Towards Cross-domain Data Analytics in Tourism: a Linked Data based Approach. *Information Technology and Tourism*, page Forthcoming (Accepted for publication), 2016. 10.1007/s40558-015-0049-5.

[44] P. E. R. 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 ICSC 2012, Palermo, Italy, September 19-21, 2012*, pages 282–292. IEEE, 2012. 10.1109/ICSC.2012.16.

[45] A. Scharl, D. Herring, W. Rafelsberger, A. Hubmann-Haidvogel, R. Kamolov, D. Fischl, M. Föls, and A. Weichselbraun. Semantic Systems and Visual Tools to Support Environmental Communication. *IEEE Systems Journal*, page Forthcoming (Accepted 31 July 2015), 2016. 10.1109/JSYST.2015.2466439.

[46] A. Scharl, A. Hubmann-Haidvogel, M. Sabou, A. Weichselbraun, and H.-P. Lang. From Web Intelligence to Knowledge Co-Creation - A Platform to Analyze and Support Stakeholder Communication. *IEEE Internet Computing*, 17(5):21–29, 2013. 10.1109/MIC.2013.59.

[47] A. Scharl, A. Weichselbraun, M. Göbel, W. Rafelsberger, and R. Kamolov. Scalable Knowledge Extraction and Visualization for Web Intelligence. In *49th Hawaii International Conference on System Sciences (HICSS-2016)*, pages 3749–3757. IEEE, 2016.

[48] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages, IEEE VL 1996*, pages 336–343. IEEE, 1996. 10.1109/MSP.2014.80.

[49] T. Tarasova. News Fact-checking: One Practical Application of Linked Statistics. In S. Capadisli, F. Cotton, A. Haller, A. Hamilton, M. Scannapieco, and R. Troncy, editors, *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference (ISWC 2014). Riva del Garda - Trentino, Italy. October 19th, 2014.*, volume 1550 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[50] 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) co-located with the 13th International Semantic Web Conference (ISWC 2014). Riva del Garda, Italy, October 20, 2014.*, volume 1279 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

[51] A. Weichselbraun, S. Gindl, and A. Scharl. Enriching Semantic Knowledge Bases for Opinion Mining in Big Data Applications. *Knowledge-Based Systems*, 69:78–86, 2014. 10.1016/j.knosys.2014.04.039.

[52] A. Weichselbraun, D. Streiff, and A. Scharl. Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *International Journal on Artificial Intelligence Tools*, 24(2):1–31, 2015. 10.1142/S0218213015400084.

[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 Co-located with 12th International Semantic Web Conference (ISWC 2013) Sydney, Australia, October 22, 2013.*, volume 1106 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[54] L. Wilkinson. *The Grammar of Graphics (Statistics and Computing)*. Statistics and Computing. Springer-Verlag New York, Secaucus, NJ, USA, 2005. 10.1007/0-387-28695-0.

[55] M. Yoshioka and N. Kando. Issues for linking geographical open data of geonames and wikipedia. In H. Takeda, Y. Qu, R. Mizoguchi, and Y. Kitamura, editors, *Semantic Technology, Second Joint International Conference, JIST 2012, Nara, Japan, December 2-4, 2012. Proceedings*, volume 7774 of *Lecture Notes in Computer Science*, pages 375–381. Springer, 2013. 10.1007/978-3-642-37996-3_32.

[56] B. Zapilko and B. Mathiak. Object Property Matching Utilizing the Overlap between Imported Ontologies. In V. Presutti, C. d'Amato, F. Gandon, M. d'Aquin, S. Staab, and A. Tordai, editors, *The Semantic Web: Trends and Challenges. 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, volume 8465 of *Lecture Notes in Computer Science*, pages 737–751. Springer International Publishing, 2014. 10.1007/978-3-319-07443-6_49.

[57] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment for Linked Data: A Survey. *Semantic Web*, 7(1):63–93, 2016. 10.3233/SW-150175.