Socioscope. A visual analytics platform for open social data

Dr. George Papastefanatos and Stavros Maroulis

Scientific Associates

Institute for the Management of Information Systems
ATHENA Research Center, Greece

{gpapas, stavmars}@imis.athena-innovation.gr

Abstract. Data visualization techniques have proven essential in Social Sciences, especially nowadays that the increasing volume and diversity of the available data make hard for people to make sense of it. The provision of interactive data exploration and visualization tools greatly assists social scientists and in general non-expert users, such as journalists, policy and opinion makers in producing value from this plethora of data. In this article, we discuss a data preparation methodology for the interactive visualization of social data and present Socioscope, a web-based visual analytics platform that allows the user to browse in a visual way various sets of social and political data related to Greece. Users can filter the data through faceted browsing and keyword search, and explore the results in a variety of visualizations for each different type of data, such as charts for numerical data, timelines for temporal data and choropleth and point maps for geographical data. Data exploration and analysis are further facilitated by such features as hierarchical representation of different levels of aggregation in charts and the ability to compare data from different datasets in the same visualization. Socioscope is presented focusing on the specific case of the PROMAP dataset, which contains data regarding the social mobilization in Greece between 1996 and 2015, and extracted via a semi-automatic way from the digital archives of two Greek newspapers.

1 INTRODUCTION

Information visualization aims at visually representing different types of data (geographic, numerical, text, network, etc.) in order to enable and reinforce cognition. Information visualization offers intuitive ways for information perception and manipulation that essentially amplify, especially for non-expert users, the overall cognitive performance of information processing. Visual analytics combines information visuali-

zation with data exploration capabilities. It enables users to explore and analyze unknown (in terms of semantics and structure) sets of information, discover hidden correlations and causalities and supports sense-making activities over data that are not always possible with traditional quantitative data analysis and mining techniques. This is of great importance, especially, nowadays that we collect and produce massive volumes of digital information concerning nearly every aspect of human activity. The so-called Big data era refers to this tremendous amount of information collected by digital means and analyzed for producing new knowledge in a plethora of scientific domains.

Social sciences traditionally depend on information processing. Data collected or extracted from a variety of sources such as surveys, focus groups, public sector information and statistics are analyzed for producing meaningful conclusions, trends and insights. Big Data greatly affects this discipline posing new challenges and problems for the social scientist. Indeed, there is a great variety of data sources (social media, the web and digital media, national statistics and research data) usually under open-access policies that provide facts that can be potentially used, combined and analyzed to explain social phenomena. However the current abundance of sources is accompanied by problems related to the large volumes and the different data types and formats (e.g. structured data in the forms of tables or spreadsheets, semi structured data or plain text from social media and web portals), the different levels of quality, trust and freshness. These characteristics make hard for the social scientist to work with large volumes of heterogeneous data without background knowledge on data management techniques for information integration, processing and querying.

Thus, nowadays the emergence of a multitude of open data initiatives around social data makes more eminent the need of having interactive human-friendly mechanisms and tools that alleviate the burden of complex data management for browsing and analysis. Tools and platforms that can offer a variety of different visualizations of the examined data, and support simple but also more complex data analysis operations in an intuitive visual manner. In this way, the actual task of social study and analysis is decoupled from the difficult part of big data processing.

Towards this goal, the S*ocioscope* (www.socioscope.gr) project aims at delivering a visual analytics platform for the social scientist to explore and analyze social facts through a user-friendly visual interface. The *Socioscope* platform offers a variety of interactive visualizations

for each different type of data, such as charts and histograms, pies and stacked diagrams for numerical data, timelines for indices and choropleth and point maps for geographical data. It is based on a multidimensional modeling approach and offers several visual operations for data exploration and analysis, such as filtering through faceted browsing, hierarchical representation of coded lists in charts, free keyword search of literal values, and capabilities for combination of different datasets along common dimensions. Moreover, it enables the reusability of knowledge by making all data available for download in various formats as well as in the form of Linked Open Data (Bizer, Heath and Berners-Lee, 2009), which is a standard means for data sharing on the web, enabling citation and unique referencing across sites.

Socioscope is used for visualizing various sets of social and political open data related to Greece, such as election results between 1996 and 2015, facts about the members of the Greek parliament, statistical data related to criminality, poverty, social inclusion, etc. Most notably, Socioscope is used for providing visualization and exploration functionality to the PROMAP dataset. The PROMAP dataset and the accompanying codebook describing the dataset's main components has been a joint project headed by social scientists of the National Center for Social Research in Greece with the participation of computational linguistics and data scientists from the ATHENA Research Center. PROMAP collected data about the protest claims between 1996 and 2015. Data related to the protest claims were extracted from articles from two Greek newspapers and were modelled according to the Promap codebook, which provides the basic classifications of the characteristics of a claim, such as the actor of the claim, the issue of the claim, the date, etc.

In this article, we present our approach, i.e., the requirements, methodology, and technology selected, for developing the Socioscope platform. We motivate our work and focus on the use case of the protest claims because its data characteristics and user requirements expressed by the social scientists pose interesting challenges regarding the variety and interactivity of the charts, the user operations for searching and exploring the claims, the different levels of granularities of the presented information. Through this use case, we present the specific requirements, modeling assumptions and visualization techniques employed as well as implementation details on the platform. Finally, we present the use case results in terms of interesting diagrams and operations the user

can perform; still the scope of these results is to present the functionality of the platform, its potential use and benefit that can offer to the social scientist, without getting into more details on possible interpretations and conclusions drawn by the study of the results.

This article first presents an overview of visualization and exploration approaches employed in social sciences, a roadmap regarding the data and visual preparation steps taken for producing interactive visualizations and the specific modeling and user requirements posed by the protest data. It then proceeds with an overview of the functionality of the Socioscope platform along with its architecture and core building elements. It then provides characteristic diagrams and operations the user can perform and explain how these data can be reused in the form of Linked Open Data for querying and linking.

2 INFORMATION VISUALIZATION IN SOCIAL SCIENCES.

Data is becoming increasingly ubiquitous in all areas of human knowledge. Social sciences can gain a lot from the plethora of open data available nowadays, be it statistical or economic, governmental and geographical data. Once using data collected primarily through such instruments as social surveys and censuses, social scientists can now benefit from the continuous stream of data from social media platforms, as well as from the development of new technologies and tools that allow the automated analysis of more traditional media of communication (e.g. newspapers).

The variety of data sources along with the volume of available data makes hard for the user, especially the non-expert, to process and extract valuable knowledge and insights out of them without appropriate tools. Visualization techniques have proven essential for helping people make sense of data and produce value from it. They may also help engage a wider audience in the exploration and analysis of social data promoting data transparency and awareness.

A key feature in data visualization is interactivity and in general the capability offered to the user for performing data analysis (searching, filtering, combining etc.) in visual ways. Allowing the user to choose from different kinds of visualization is also crucial, since no single visualization configuration suits every data analysis context. For example,

map-related visualizations, such as choropleth maps and heatmaps are suitable for geographical data, network data are usually represented with graph-related visualizations, whereas statistical data and indices may be better visualized via traditional charts, such as line and area charts, timelines, pies, stacked diagrams, and scatter plots. The visual Information-Seeking Mantra (Shneiderman, 1996) summarizes many visual design guidelines and provides an excellent framework for designing information visualization applications. The core concept is: "overview first, zoom and filter, then details on demand", which illustrates the importance of gaining an overview of the data and then interactively exploring a more detailed representation in the visual analysis scenario.

In recent years, online tools have been introduced that offer access to data through interactive visualizations. Data providers themselves, usually offer ways to browse their data in visual ways. For example, Eurostat, the statistical office of the European Union, offers several tools for the visualization of the data it provides. Some of them are centered on a specific theme, offering infographics and visualizations to provide insight into it, like "My Country in a Bubble"[1], which compares EU countries along several indicators with a bubble chart, or "Young Europeans"[2], which allows users to compare themselves with other young men and women in the EU area. Other Eurostat tools, like the Eurostat Table, Graphs and Maps Interface (TGM), offer a more general way of exploring the Eurostat datasets through several visualization types such as maps, bar, line or pie charts and scatter plots. Data aggregators, like Google Public Data Explorer[3] or Gapminder[4], collect data from various open data providers and make it available for users to explore and compare interactively. Google Public Data Explorer hosts a range of public data from several organizations like the U.S. Census Bureau, Eurostat, World Bank, etc. A user can also upload her own datasets to Google Public Data Explorer for visualization and exploration. Gapminder is a tool that allows the user to visualize the evolution of multiple indicators and statistics about social, economic and environmental development at local, national and global levels.

[1] http://ec.europa.eu/eurostat/cache/BubbleChart/
[2] http://ec.europa.eu/eurostat/cache/infographs/youth/index_en.html
[3] http://www.google.com/publicdata/directory
[4] http://www.gapminder.org/

3 A ROADMAP FOR VISUALIZING AND EXPLORING SOCIAL DATA

In this section, we describe the requirements and a generic methodology for producing meaningful visualizations for social data and then we focus on the specific steps taken for processing the protest data. The goal is to offer a user-driven end-to-end roadmap of the necessary tasks, problems and challenges related to the visualization of social data.

The overall process is shown in **Fig. 1**. The source information is usually present in various formats according to each source and the appropriate data extraction and analysis technique must be selected, such that raw information is transformed to a more semantically-rich, structured format. Then, a set of data processing techniques are applied to enhance the quality of the collected data, such as cleaning data inconsistencies, filling in missing values, detecting and eliminating duplicates, etc. Moreover, the data is enriched and customized with visual characteristics, meaningful aggregations and summaries are produced for enabling the user-friendly data exploration and proper indexing is performed for enabling efficient searching and retrieval. Next we present the details for each step.

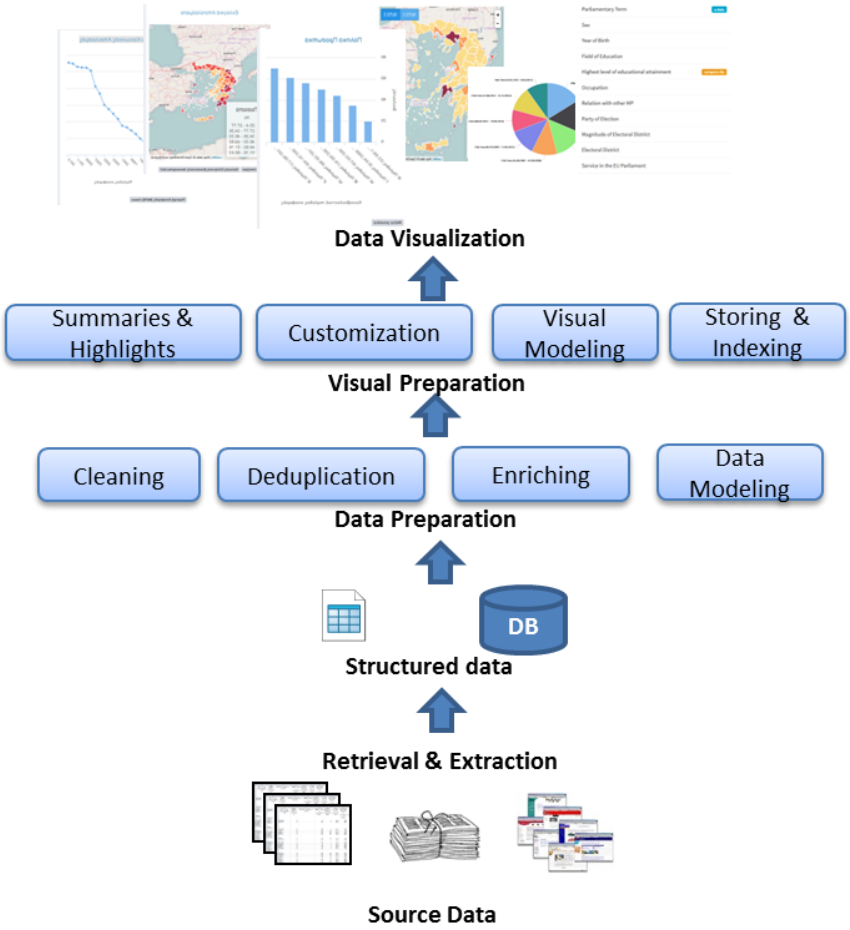

Fig. 1. Steps for visualization of social data

Data retrieval, extraction. The first step concerns the retrieval and extraction of the data to be visualized. Raw data exists in various different formats; newspapers, books and reports describe social phenomena in plain text, web sites and social networks contain annotated text and semi structured data whereas surveys, open data sources and web databases provide structured information. Data must be first retrieved in an appropriate for processing digital format (e.g., digital text files from newspapers), out of which the core modeling concepts and observations are extracted. This task is usually performed in an iterative human-

curated way that refines the quality of the extracted data, especially when the source data is in plain text. For structured data sources, this process involves the extraction and mapping of the source concepts to the target modeling.

Data Preparation. Input data are provided either in databases or data files (.csv, .xml, etc.). This step involves identifying all the concepts within the input datasets and representing them in a uniform data model that supports their proper visualization and visual exploration. For example, as described in details in section 3, Socioscope uses the multidimensional model for representing concepts as observations and dimensions. Observations are measures of social phenomena (e.g., indices) and dimensions are properties of these observations (e.g., reference period, reference area, etc.). Thus, a first step is to analyze the datasets and identify the different types of attributes contained (date, geolocation, numeric, coded lists, literal). Each attribute is mapped with the corresponding concept of the multidimensional model, such as dimension, observation, coded list, etc. In addition, data processing requires a set of quality improvement activities that eliminate data inconsistencies and violations in source data. For example, missing or inconsistent codes are filled for coded list attributes, date and time attributes are transformed to the appropriate format, and numerical values are validated such that wrong values are corrected. Moreover, input data coming from multiple sources usually contain duplicate facts and deduplication of the dataset must be performed. Deduplication is the process of identifying duplicate concepts within the input dataset based on a set of distinctive characteristics (e.g., regarding the protest data, this set can contain the protest claim title, the reference period and the location of the claim). A final task concerns the enrichment of the data with information coming from external sources. For example, places and locations are usually extracted as text; they can be annotated and enriched with spatial information (e.g. coordinates, boundaries, etc.) from external web services for their proper representation in maps.

Visual Preparation. This set of tasks involves the enrichment and customization of the data with characteristics that enable the proper visualization of the underlying information. These characteristics extend the underlying data model with visual information. For example, colors can be assigned to coded values and different types of diagrams can be bound to different types of data (timeline to date attributes and maps to geographical values). Thus, customization and building the

visual model is a necessary step for data visualization. In addition, the production of visual summaries and highlights is a common task especially for the visualization of very large datasets. Summaries provide to the user overviews of the visualized data and visual highlights are used to present interesting charts and findings. Finally, all data are stored and indexed in an appropriate format such it can enable efficient exploration and searching. For example, traditional RDBMS systems can be used for exploration and visualization of tabular data, novel NoSQL database systems, such RDF and Graph databases can be used for visualizing network and graph data and inverted indexes can be used to support text search capabilities.

Data Visualization. The final task on this process is the actual visualization of the data. This involves the provision of the different types of charts, maps, and graphs that present the data and the visual ways, such as search, browsing, filtering, etc., for performing data analysis. Different types of charts can be offered according to the type of information; numerical and tabular data can be presented through typical charts, such as bar and line diagrams, pies, stacked and scatter diagrams, areas, etc. Temporal (dates and time periods) information can be visualized with timeline diagrams. Hierarchical information (e.g. hierarchical coded lists) can be explored with hierarchical diagrams such as tree maps and nested diagrams. Network data are visualized as graph diagrams providing explicit representation of the interrelationships of the visualized objects. Finally, geographic information is usually visualized on maps; choropleth, heatmaps and bubble maps capture the density of an observation over a region, point and clusters can be used for presenting the location of individual entities on maps. Multiple charts can also be combined and offer more sophisticated visualizations to the user.

Regarding the visual analytics capabilities, gaining overview, filtering out, specializing and presenting results in multiple ways are crucial tasks in the visual exploration scenario. Thus, data visualization tools can provide dashboard-like overviews and catalogs of the available datasets and highlights of the most interesting ones, faceted browsing for applying filters to the data, free keyword search for finding text in literal values, zoom-in and zoom-out operations for exploring data at different granularity levels (geographic regions, date periods, etc.). Also, they should offer the ability to view multiple series and compare different datasets into joint diagrams. Last but not least, tools must enable

reusability of data; all data must be available for download under an open access license and follow the best practices for publishing open data on the web.

3.1 The case of Protest Data

The main entity the protest data is the notion of the claim. A claim refers to an expression of a political opinion in the public sphere, taking the form of a physical or verbal action, and is composed of several components, such as the actor who poses the claim, the addressee of the claim, etc. In the following section, we present in more details the aforementioned roadmap for the case of the protest data.

Regarding the data extraction and preparation, all claims were extracted via a semi-automatic way from the digital archives of two Greek newspapers between 1996 and 2014; articles from two specific days of the week have been selected for processing. The free text was parsed and transformed to semi-structured annotated text, where the social scientists validated the entity and the characteristics of a claim. Each of the characteristics of a claim was represented by a number of variables following the proposed Codebook. A date variable (CDATE) is used to denote the date the claim was made. The actors making the claim are specified by a free-text description (ACT), as well as by a coded variable (ACTS) that refers to the more general category the actor belongs to (Tertiary Trade Unions, Students etc.). Similar variables (ADR, ADRS) are used to represent the addressee of the claim that refers to the actor at whom the claim is addressed. The claim's issues, i.e. what the claim is about, are also described by a free-text field (ISSUE) and categorized with the coded variable ISFIELD, which takes values like "Wage policies, employment and social benefits", "Working conditions and labor rights", "Education" etc. The form with which the claim was expressed in the public sphere is specified in two levels of detail with the coded variables FORM (Strike, Occupation/sit-in, Blockade etc.) and FORMS (Labor related protests, Demonstrative protests), the latter of which summarizes the former. Finally, a claim may be associated with a specific geographic location.

The data was initially stored and handed to us in a spreadsheet format. There was a sheet containing the primary data, as well as a sheet for every coded dataset variable, each of which listed an id and a human-readable label for every possible value of the variable it described.

In the primary sheet, every single row represented a single claim, containing values for its various fields. For coded variables, the id of the corresponding code was used. The rest of the row cells contained the non-coded dataset's fields (free-text, number or dates).

At first, we had to identify the various components the data consisted of and map them to the data model used in Socioscope. Initial data contained location information in textual form. To visually represent claims as points on a map, we had to translate this text to geospatial coordinates using a geocoding service. Next, we enriched the data with properties needed to properly visualize it. Every field was annotated with metadata that determined the types of visualizations suitable for it, as well as if it would be used for faceting. For coded variables, we also added a property that denotes the default order by which its values should be presented in the x-axis of a chart, while for temporal variables, a property was used to denote the different date intervals by which it could be visualized in a timeline (e.g. year, month).

For the visualization of the protest data, the main requirements were to allow the user to get an overview of the data through meaningful charts and maps, while still retaining access to the detailed representation of each claim. Additionally, since protest data contained free text fields, we wanted the user to be able to perform full-text search through the data and visualize the results. To satisfy these requirements, after transforming the data into a suitable format (JSON), we indexed and stored it in a full-text search server (ElasticSearch).

4 SOCIOSCOPE OVERVIEW

Socioscope is a platform that combines faceted browsing with dynamic generation of visualizations, in order to provide a user-friendly data exploration experience. Its main features are:

- **Dashboards and Highlighted visualizations.** Socioscope offers a dashboard-like visual experience that enables users to gain an overview of the underlying datasets through visual highlights. For each dataset, a number of different visualizations and summarizations can be defined and presented at the home page. Such visualizations are meant to give an insight of the data contained in the dataset and offer an interesting view of it.
- **Faceted dataset exploration.** User exploration is performed via an intuitive faceted browsing way, such that filtering and specialization over the data does not require a priori knowledge of the structure or all possible values of the dataset's attributes. Faceted browsing presents a list with the dimensions of the dataset along with the distinct values of each dimension. The user filters the visualized data, by selecting one or more of the presented values. In this way, the user can apply multiple filters on different dimensions and set the x-axis (e.g., for charts) to a specific dimension.
- **Multiple Chart Generation.** Socioscope supports a rich variety of chart types (line, bar, column, area, pie) enabling interactive statistical analysis. A set of configuration options are available for the charts, such as show or hide measure values in a chart, enable chart stacking(normal or percentage), as well as change the granularity of a time axis (e.g. year, month or day intervals).
- **Hierarchical Charts.** A dimension's code list may have a hierarchical structure in order to organize data at different levels of aggregation. For example, the administrative regions may follow a NUTS classification organizing municipalities into regions, prefectures, etc. Socioscope identifies the hierarchies of such dimensions and offers a hierarchical browsing functionality to the user who can easily zoom-in and out to the results. Initially, the top level of the hierarchy is presented, providing an overview of the data. Transitions between different levels of the hierarchy are performed through user interaction.

- **Choropleth, Point and Cluster Maps.** Socioscope offers a variety of different map representations for visualizing dataset containing geographical information. For example, a measurement that varies across a geographic region can be represented on a choropleth map where different areas are colored in proportion to the density, i.e., value of the measurement. The choropleth map provides a visual way for the variation of a measurement over a geographic area. Also, Socioscope offers point and cluster maps, in which geospatial information represented by coordinates are drawn as points at their corresponding geographic location. Zooming out on a region of a point map groups together individual points, eventually forming clusters on the map that contain statistical information regarding the underlying clustered data. This functionality enables users to easily go from coarse-grained overviews to individual points on the map with only a few mouse clicks.
- **Search Functionality.** Socioscope provides full-text search capabilities, which are particularly useful for filtering through data fields that do not follow a coding scheme, but contain free text instead. By performing a full-text search, users can visualize data that contain the words they searched for.
- **Browse individual data entities.** Data visualization in Socioscope does not only involve aggregated statistics and summarizations. Through the faceted browsing interface used for generating data visualizations, users can also browse through the raw data entities, which are presented in a user-friendly tabular way.
- **Combine data from different datasets.** Data from different datasets can be combined in the same visualizations, and provide interesting correlations and causalities. The combination of different datasets can be achieved in datasets that share at least one common dimension or a dimension that contains the same kind of values (e.g. datetime values).
- **Visualization exports and datasets download.** Users can export an interesting chart and download it in various formats (e.g. PDF, JPG, PNG), as well as print it. Also, all datasets are publicly available and can be downloaded as RDF or CSV dumps for further processing by the users .
- **API availability and Query Functionality.** Finally, a RESTful API service is provided through which users can execute arbitrary queries and browse the raw data.

4.1 Socioscope Data & Visualization model

The underlining data model behind Socioscope, is an adaptation of the more generic Multi-dimensional Data Model, and specifically the RDF Data Cube Vocabulary (Cyganiak, Reynolds and Tennison, 2014). The Data Cube Vocabulary is used for the description and publication of multi-dimensional statistical data. A statistical data set comprises a collection of observations made at some points across some logical space. A set of dimensions indicate what an observation applies to (e.g. time, area), whereas measures indicate the phenomenon being observed. An observation can be further interpreted by a set of attributes (e.g. units of measure).

Most of the datasets in Socioscope, are statistical in nature and follow tightly the aforementioned model. However, in the case of protest data, we had to deal with non-aggregated raw data, where each entity described a single claim, with values set for its various fields. Some of the visualizations that were to be supported (e.g. the evolution of the number of claims through time) require that we perform aggregate transformations of the data, effectively creating dataset 'observations' on the fly.

At the application level, the main entity modelled is the Dataset. A Dataset represents a collection of similar data, described by a set of fields as well as aggregate measures to be calculated over it (e.g. number of claims). Most metadata components are uniquely identified by an id value, and contain a human-readable label. A dataset is also characterized by the following metadata fields:

- a short description
- the subject/category that it belongs to
- the date it was first added
- the last modification date
- a url to the original source of the data

A dataset's field is further described by its type which can be one of the following:

- coded field where its value comes from a predefined list of values
- free-text field
- date field

- geolocation field which contains geographic coordinates to a specific location

Finally, a measure is defined by its unit of measure, e.g., the population measure is identified by the number of people, etc.

Visual Model. Generating dynamic visualizations on the basis of user interaction requires that we model various characteristics that control the user operations and the visual results. These characteristics amend the underlying data model with properties, which are used for holding visual information associated with each dataset or specific entities within a dataset as well as user actions and preferences associated with visual operations on the data.

The core entity of the visual model is the visualization component encapsulating the data to be visualized along with the visualization type (e.g., chart, map, facets, tabular representation) and its parameters that configure the way information is presented to the user.

The visual model holds only the part of the underlying data that is requested or filtered in each user action. This information is modeled in the form of a series. A series is a set of data points, i.e., a set of (x, y) pairs. A data point relates a numeric measurement (y-axis) to a value of a dataset's dimension (x-axis). When the user requests a specific visualization, the visual model assigns the x-axis and y-axis attributes to a dimension and a measure, respectively. Multiple series in x-axis are constructed and modeled in the same way.

Other parameters include the visualization type and its parameters. The model stores the type of the chart selected for visualizing the dataset (e.g. line chart, choropleth map). Also it stores whether chart stacking is enabled. A stacked chart is useful in charts with multiple series, where it emphasizes the relationship of individual data points to the whole. The visual model holds information for individual attributes as well as attributes with date values. In this case, the model holds the granularity of the date interval over which data is aggregated (e.g. year, month). Also, it holds the ordering (ascending or descending) of the values composing the x-axis.

Another piece of information captured by the visual model involves the filters set by the user. The user can add a filter on a coded field, on a date field in the form of a date range (e.g., from - to), or a search filter in the form of a set of keywords. Filters are used for restricting the space of the visualized results and can be used in combination with

each other (e.g., a keyword filter can be applied together with a coded filter).

Finally, comparing data by another dataset's dimension/field is possible by setting another field as the visualization's split-by parameter. With this parameter, data is organized as a set of multiple series, one for every value selected by the user for that field.

4.2 Socioscope Architecture & Implementation

The architecture of Socioscope is presented in **Fig. 2**. Socioscope is a web-based platform with 3 primary components in its client-server architecture: the data storage layer, the user front-end, and the back-end that processes all users' requests and implements the application business logic. Next, we describe the basic components of Socioscope.

Data Storage. The data storage layer stores all data and metadata. It communicates with the back-end service in order to allow the retrieval and processing of the data served to the visualization tool. All data are stored in RDF format in OpenLink Virtuoso[5], which is an open source RDF database. As described in mire details in section 5, RDF formats are popular for representing semantically rich information on the web in the form of open data. In addition, we have used Elasticsearch[6], a full-text search server, for querying and aggregating on the fly, non-aggregated data and enabling keyword search on the data.

Back-end Data Api. The back-end is composed of a Data Access Layer that abstracts the logic necessary to access the database and collects all the necessary dataset data and metadata. The Facets Generator generates facets over a dataset's fields, while the Series Generator creates the requested series objects. In order to improve the application's performance and responsiveness as perceived by the end-user, a caching component is used. The back-end of our platform was developed in Java using the Spring Framework. Jena framework is used for handing of RDF data.

Client Application. The visualization tool at the front-end was developed as a single-page Javascript web application, using the Backbone.js framework. In a single-page application the presentation logic is mostly handled in the browser, while the back-end is responsible for

[5] http://virtuoso.openlinksw.com/
[6] https://www.elastic.co/

supplying all the required data. In an effort to provide an optimal browsing experience across a wide range of devices (e.g. smartphones, tablets), we have developed the web user interface following the responsive web design approach.

The Visualization Controller component, based on user interaction, fetches the appropriate data from the back-end Data Api and instantiates the appropriate client-side data and visual model entities. Depending on the visualization type selected, the corresponding visualization generation module (Charts Generator, Maps Generator and Search Results Browser) prepares and displays the resulting visualization. Regarding visualization libraries, we use Highcharts for the creation of the various supported charts and Leaflet.js for the map visualizations.

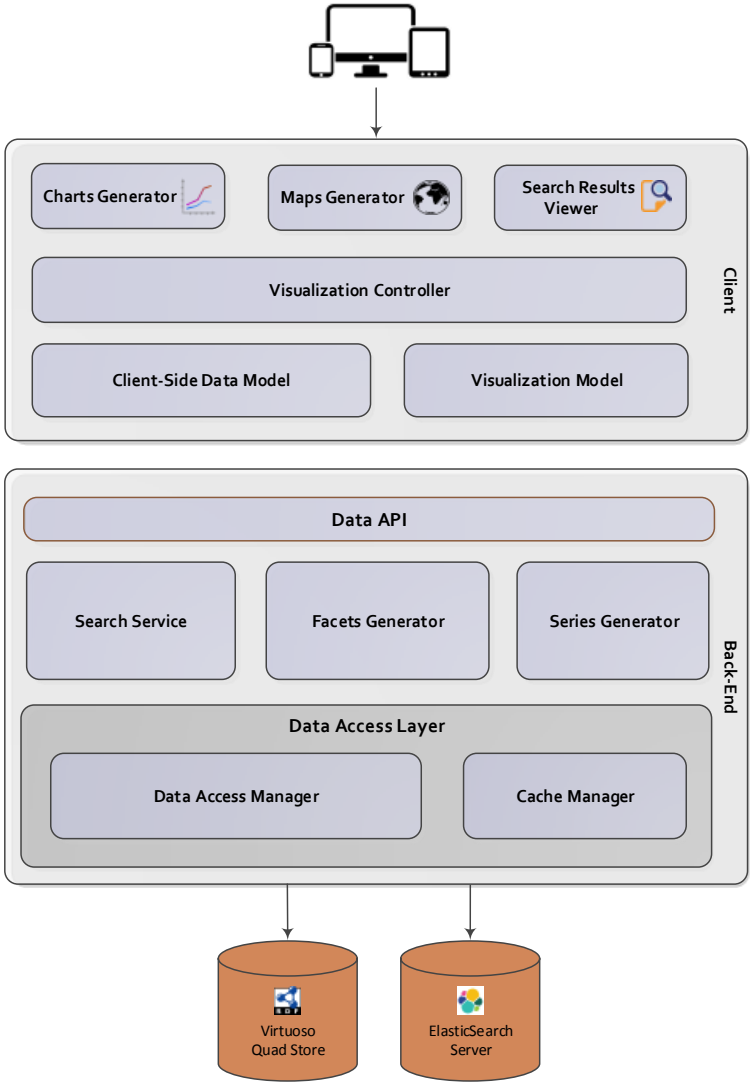

Fig. 2. Socioscope Architecture

4.3 Socioscope Presentation

In this section, we outline the use case of protest data scenario with the goal of exemplifying the visualization capabilities of Socioscope. In the home page of Socioscope, users have access to a dashboard with the

highlights of available datasets, from which they are able to preview and select the dataset they wish to explore (**Fig. 3**).

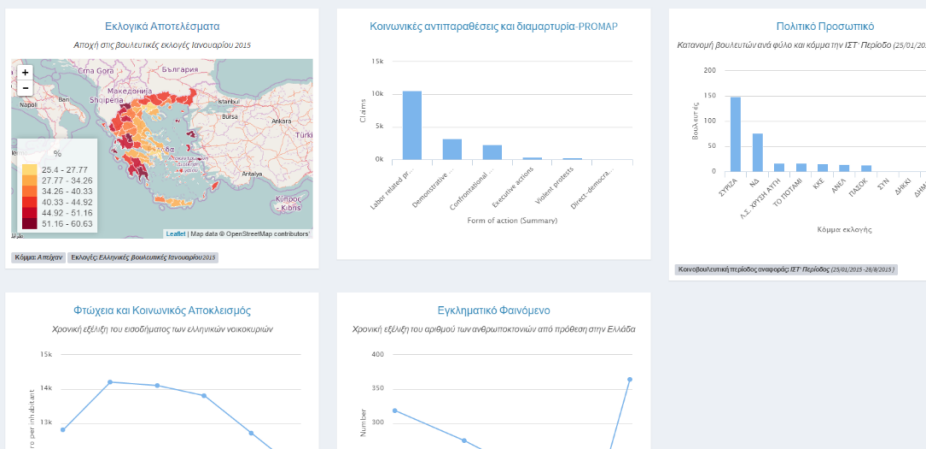

Fig. 3. Dashboard-like view of dataset highlights

For each dataset, the primary data exploration and visualization user interface appears, which consists of two main panels:

- A Facets panel (**Fig. 4**) that enables the faceted browsing over a dataset and that lists all the dimensions and values contained in the dataset. The Facets panel offers a set of visual operations that can be used by the user to filter the results shown, set the x-axis and y-axis of the chart, include multiple series and visually compare data from other datasets. A text box is also included, which is used for performing full-text search within the dataset.
- The Visualization panel which is the area where the generated visualization is presented. At the top of the visualization panel, there exist graphical control elements that allow the user to change the type of the visualization being displayed, customize it and export it in various formats.

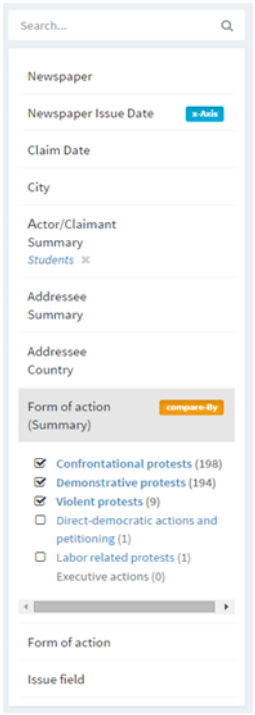

Fig. 4. Facets Panel

Fig. 5 shows a timeline of the protests acted by Students that have been characterized as Demonstrative, Confrontational or Violent. As can be seen in the filters panel shown in field "Newspaper Issue Date" has been set on the x-axis of the chart. That is a date field, and as a result, the chart generated is a time series graph presenting the total number of claims per newspaper issue year. Enabling the "compare-By" option on the field "Form of Action (Summary)" and checking on these 3 available options, results in multiple series being drawn on the chart, one for every option checked. A filter has been added to the field "Actor/Claimant Summary", which means that the chart shows the number of claims that were made by "Students".

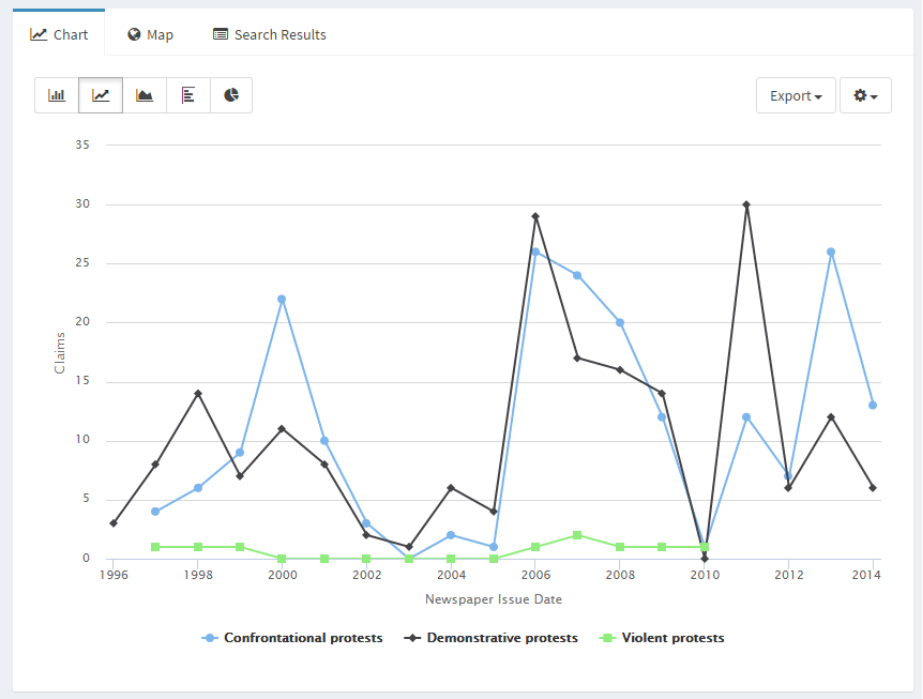

Fig. 5. Timeline of student protest claims

Fig. 6 shows the same data depicted in the timeline of **Fig. 5**, but as a stacked column chart.

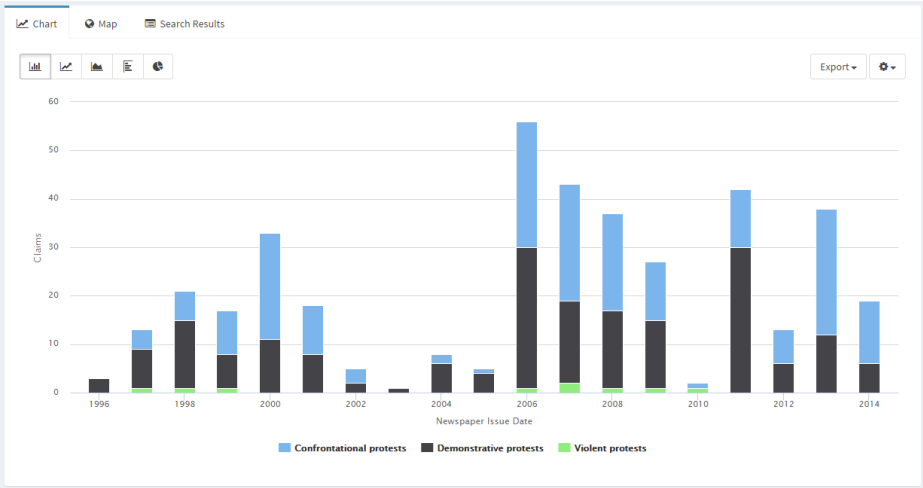

Fig. 6. Stacked column chart of student protest claims

A Claim may contain geographic information in the form of geographic coordinates. Using this information, claims satisfying the filters added by the user are depicted on a map. For example, the map in **Fig. 7** displays claims referring to farmers' blockades.

As it is shown, farmers' blockades are visually depicted in clusters on major regions in Greece. Instead of plotting every marker on the map, which may lead to visual overload, multiple neighboring markers are clustered and each resulting cluster is drawn as a single marker displaying the number of claims it contains. Zooming in, reclusters the markers, showing a more detailed view, while zooming out leads to even less markers getting drawn. Clicking on a single marker results in a pop-up box being displayed that contains detailed information for the claim it refers to. In this way, a point map allows a user to identify areas of particular interest (e.g. areas with high density of data points), zoom to them and explore the corresponding data in full detail.

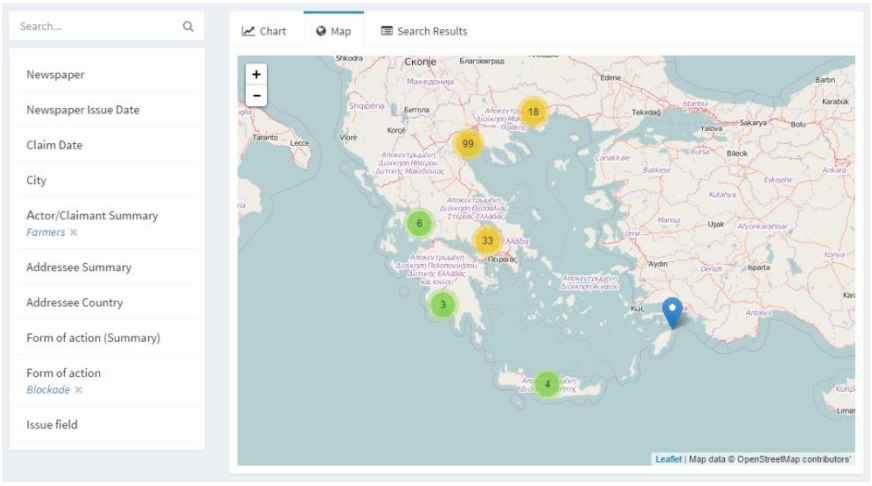

Fig. 7. Map depiction of farmers' blockades

An example of this is shown in **Fig. 8** which contains the same map visualization as above, but zoomed in to get a much more precise view. It also shows the pop-up box that appears after clicking one of the markers, and specifically, a marker that refers to a farmers' blockade on a national road of Greece in Thessaly. From that box, it can be seen that the specific blockade was accounted in January 2002 in one of the two Greek newspapers used for the data extraction. The headline of the

newspaper article from which the claim was extracted is also included, allowing the user to further interpret it.

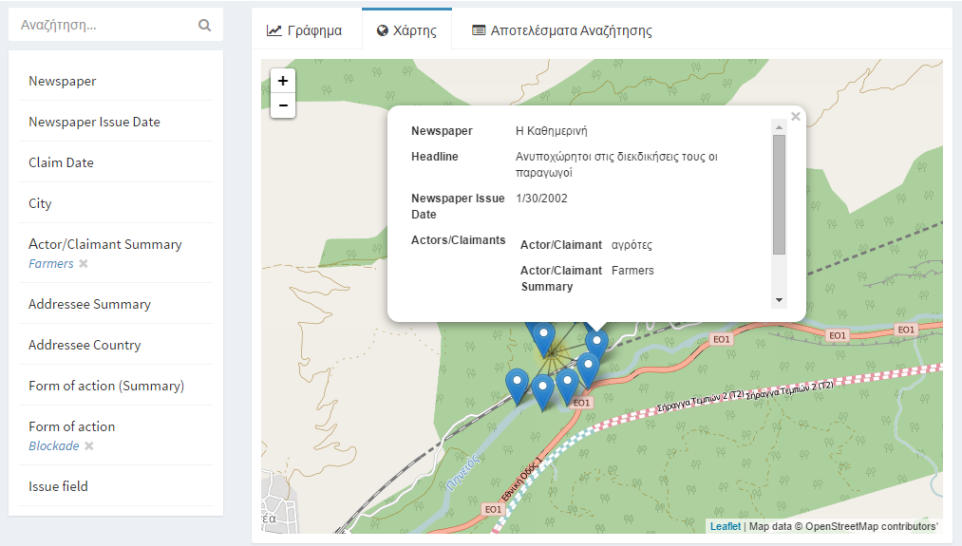

Fig. 8. Zoomed-in point map

Switching to the Search Results Tab, shown in Fig. 9., displays a list of the actual data entities that satisfy the filters added. This tabular view contains detailed information for each claim (i.e., raw data) such as the newspaper used as its source, its headline, the date of the newspaper's article as well the date of the claim, the actors and the forms of actions, etc.

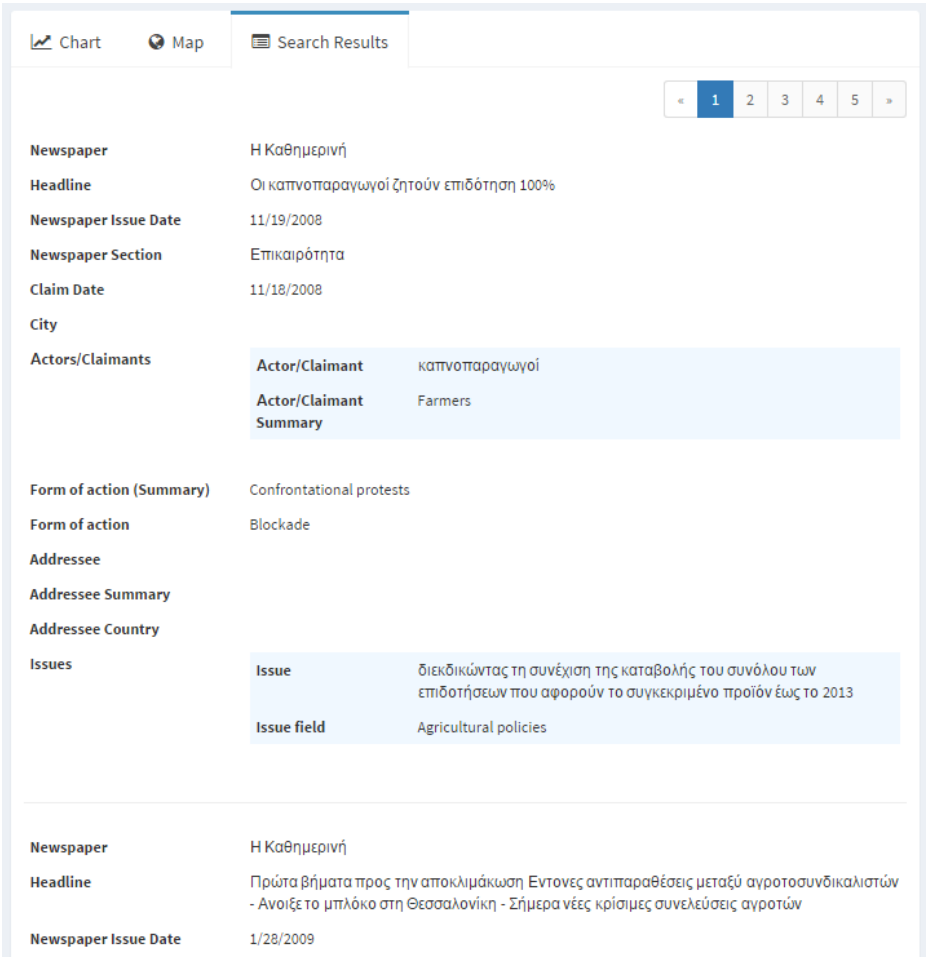

Fig. 9. Tabular View

Fig. 10 and **Fig. 11** illustrate how hierarchically organized dimensions are visualized in a chart, with data shown initially at a higher level of aggregation, and more detailed views presented through user interaction. Specifically, in the chart shown in **Fig. 10**, the x-axis represents the second level of the NUTS classification scheme for Greece. After clicking on the column that refers to the region Crete (Kriti) in the chart, the user is presented with data for the regional units that Crete is divided into (NUTS 3), shown in **Fig. 11**.

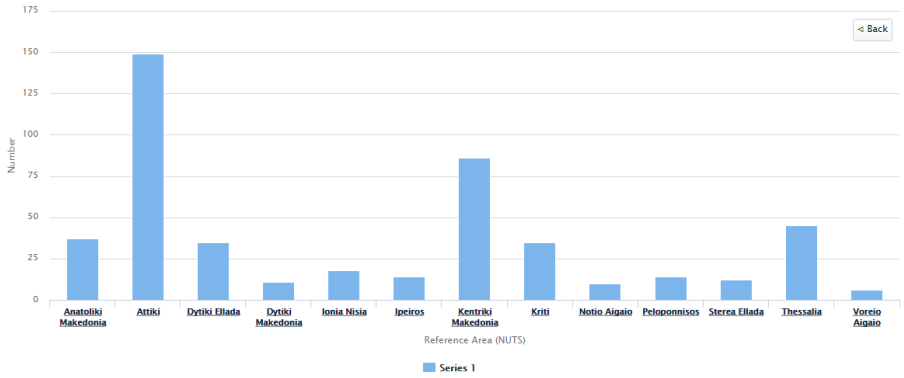

Fig. 10. Hierarchical Chart

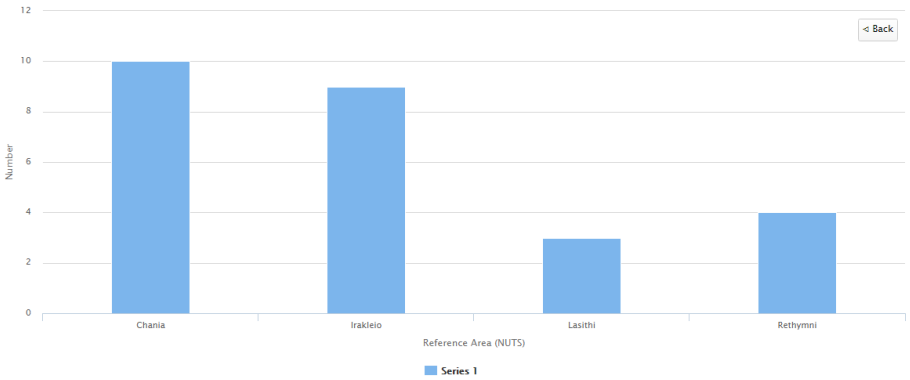

Fig. 11. Regional units of Crete (Kriti) region after zooming on Fig 10.

5 SOCIAL DATA AVAILABLE AS LINKED OPEN DATA

The recent exponential growth of the data on the web and the plethora of available content have changed the way we create, interlink and consume information. More and more governmental, scholar and user-generated datasets are published with open access, and become available for potential data consumers. The Linked Data paradigm (Bizer, Heath and Berners-Lee, 2009) is a common practice for publishing, managing and sharing structured information on the Web, and offers new ways of data integration and interoperability. The main concept in

Linked Data is that all information resources (e.g., a dataset, a classification and its values, a person or a concept, a location, etc.) published on the web are uniquely identified by a URI, and typed links between URIs are used to connect resources. The modeling approach used for representing Link Open data is RDF[7], a W3C standard model for data interchange on the Web, and SPARQL[8] is the language for accessing and querying such data.

In **Fig. 12**, we demonstrate a simple example of a claim acted by Students with the form of a Blockade. The concept of a Blockade (form of action of a claim) is uniquely identified by a URI: http://www.socioscope.gr/terms/Blockade, and the concept of Students (being the actor of a claim) has the URI http://www.socioscope.gr/terms/Students. A claim is also assigned with a unique URI, http://www.socioscope.gr/reource/claims/12456 and is connected with the two aforementioned concepts with two typed links (http://www.socioscope.gr/hasActor and http://www.socioscope.gr/hasFormOfAction), also being identified via URIs. In this manner, data are represented as interconnected resources and published on the web in the form of URIs and links between them.

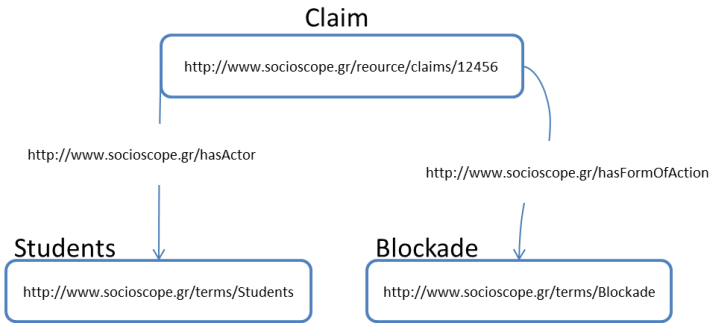

Fig. 12. A Linked Data representation of a claim in RDF

Reusing existing URIs across multiple open data sites rather than creating new ones, and pointing from one dataset to another by referencing these URIs, forms the Linked Open Data cloud.

Socioscope uses the methodology described in (Petrou, 2014) and (Petrou, 2013) for converting all data in RDF format and making it

[7] https://www.w3.org/RDF/
[8] https://www.w3.org/TR/rdf-sparql-query/

available in the form of Linked Open Data. Data is available via two forms: RDF dumps, e.g., files that can be downloaded and processed locally or via a SPARQL web API where users (humans or other web services) can pose queries on the data and get results in HTML format. The gain of making Socioscope data available as Linked Data is two-fold: First it offers a uniform way of data representation on the web enabling the interlinking and reuse of social information from other open-access social data sources. This makes the data comparable across multiple sites, fostering social research at the web scale. Second it maximizes the dissemination and the citation of the concepts and results presented within the platform, as Socioscope information is uniquely identifiable from other sources at a most granular level, e.g., each individual claim is assigned with a unique id (URI), a URI can link to a collection of claims fulfilling a set of filters, or most importantly URIs are assigned to concepts and values contained within the Promap Codebook.

6 CONCLUSIONS

In this article, we have analyzed an approach to interactively visualize social data. Visual analytics tools have proven essential in helping people make use of data, especially nowadays that data is collected in greater volumes than ever and that the open data movement has made a considerable amount of it available to the general public. All this data is not in the same form and quality differs between various sources, making it very important to streamline the steps needed for the preparation of data for visualization. Socioscope, a platform which makes it possible to explore several sets of social and political data related to Greece, provides a web-based user interface that allows the user to dynamically create various visualizations. Since no single visualization type suits every data exploration scenario, Socioscope offers an array of different types of charts and maps, as well as a tabular view for the representation of raw data entities. An intuitive faceted browsing and keyword search interface enables the filtering of the data visualized and features like hierarchical charts and marker clustering in point maps allow the user to interactively represent data in different levels of details. The PROMAP dataset proved challenging in regards to its visualization requirements, and thus, we have presented our approach focusing on this particular use case.

Socioscope is a work in progress, and we plan to further extend it with new features such as additional interactive visualizations or the ability for a user to visualize her own external data and compare it with the datasets hosted in the platform. Also, we developed it with the idea of effortlessly enriching it with new datasets in the future. To this effect, we employed standardized technologies and vocabularies like the RDF Linked Data format and the Data Cube Vocabulary that allow the reuse of such metadata resources as dimension code lists, measures and administrative ontologies. Besides, more and more social data are published following the Open Linked Data paradigm.

7 REFERENCES

1. Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pp. 205-227.
2. Cyganiak, R., Reynolds, D. and Tennison, J. (2014). The RDF Data Cube Vocabulary, World Wide Web Consortium. Available from: http://www.w3.org/TR/vocab-data-cube/
3. Petrou, I., Meimaris, M., and Papastefanatos, G**.** Towards a methodology for publishing Linked Open Statistical Data. In *eJournal of eDemocracy and Open Government 6(1): 97-105*, December 2014.
4. Petrou, I., Papastefanatos, G., Dalamagas, T. Publishing census as linked open data: a case study. *In 2nd International Workshop on Open Data (WOD 2013)* – Paris, France, June 2013. 4:1-4:3.
5. Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings of the IEEE Symposium on Visual Languages*, pp. 336-343, Washington, D.C.: IEEE Computer Society Press.