Automated Visualization Support for Linked Research Data

Belgin Mutlu[1], Patrick Hoefler[1], Vedran Sabol[1], Gerwald Tschinkel[1], and Michael Granitzer[2]

[1] Know-Center, Graz, Austria
[2] University of Passau, Germany
{bmutlu,phoefler,vsabol,gtschinkel}@know-center.at
michael.granitzer@uni-passau.de

Abstract. Finding, organizing and analyzing research data (i.e. publications) published in various digital libraries are often tedious tasks. Each digital library deploys their own meta-model and technology to query and analyze the knowledge (in further text, scientific facts) contained in research publications. The goal of the EU-funded research project CODE is to provide methods for federated querying and analysis of such data. To achieve this, the CODE project offers a platform, that extracts scientific facts from research data and integrates them within the Linked Data Cloud using a common vocabulary (i.e. meta-model). To support users in analyzing scientific facts, the project provides means for easy-to-use visual analysis. In this paper, we present the web-based CODE Visualization Wizard, which aims to analyze research data visually with an emphasis on automating the visualization process. The main focus of the paper lies on a mapping strategy, which integrates various vocabularies to facilitate the automated visualization process.

Keywords: Linked Data; Visualization; Research Data; RDF Data Cube

1 Introduction

Digital libraries, which control the lifecycle of research publications (i.e. publishing and making them accessible for certain communities) mainly expose the research knowledge using domain-specific meta-models and technologies. Moreover, they only focus on some structural attributes and often don't consider the content of the publications. This domain-specificity and weakness in specifying querying attributes limit the ability to effectively find desired information, since the number of published content is continuously growing. The goal of the CODE[1] [4] [5] project is to offer a solution for this issue by providing a platform that structures (heterogeneous) research data using the RDF Data Cube Vocabulary[2] and releases them as Linked Data.

[1] CODE: http://code-research.eu/
[2] RDF Data Cube Vocabulary: http://www.w3.org/TR/vocab-data-cube/

The RDF Data Cube Vocabulary is a generic vocabulary used to describe quantitative data (e.g. research results from tables). To simplify the analysis of this data, the web-based CODE Visualization Wizard[3] has been developed, which integrates several visualizations. To achieve a Linked Data-based visualization, these visualizations (e.g. charts) should also be described semantically. For this purpose, we defined the Visual Analytics (VA) Vocabulary[4] in the form of an OWL ontology. This vocabulary is an interface between the RDF Data Cube and visualization-specific technologies, and together with the RDF Data Cube Vocabulary it forms the basis for automating the visualization process.

In this paper, we summarize the current status of the CODE Visualization Wizard and its ongoing research.

2 Related Work

Semantic description of visualizations using RDF is a new research topic and the literature, up to now, offers just a few related publications. The most significant research, the Statistical Graph Ontology [3], comes from the biomedical domain and presents a new approach to annotate visualizations semantically.

While the Statistical Graph Ontology provides a sophisticated ground for describing statistical graphs, some key issues (e.g. the description of size and color as visualization component or the datatype of a visualization component etc.) for our applications were missing. This is why we have extended this vocabulary for our Visualization Wizard.

At Stanford University, an interactive Web-based visualization system, the Vispedia [1], has been developed to visualize heterogeneous datasets. The visualization process of Vispedia is based on the integration of the selected data into an iterative and interactive data exploration and analysis process enabling non-experts to more effectively visualize the semi-structured data available. Vispedia was an inspiration for the Visualization Wizard, but being a Wikipedia plugin it only supports visualization of Wikipedia data. Also, it does not provide automatic binding of heterogeneous data onto visualizations.

3 Approach for Automated Visualization Support

In contrast to other available solutions for visualizing Linked Data [2], the CODE Visualization Wizard automatically suggests suitable visualizations based on (1) the content and structure of the provided research data and (2) semantic description of the visualizations. The following parts of our wizard contribute to these features:

Vocabularies: The RDF Data Cube is a W3C Standard and has been developed to represent statistical data as RDF. In the CODE project we use this standard to define the meta-model for the basic research data in order to capture the

[3]CODE Visualization Wizard: `http://code.know-center.tugraz.at/vis`

[4]VA Vocabulary: `http://code-research.eu/ontology/visual-analytics`

evaluation results from publications. The results are represented as a collection of observations consisting of a set of dimensions and measures, which represent the structure of the data. Dimensions identify the observation, measures are related to concrete values and attributes add semantics to them. For example: when we have a dataset representing the result of a scientific challenge (such as PAN[5]) for several teams, there will be a collection of observations with dimensions describing the teams with concrete values for the challenge result and with an attribute *percent* to identify the unit of the value it is measured in.

Our VA Vocabulary is used to represent the information about visualizations. It describes the visualization axes and other visual channels, such as color or size of visual symbols, used to visually represent the data. The vocabulary also describes suitable datatypes that can be represented by the axes and visual channels, including the allowed occurrence of the axes and visual channels. The definition of the occurrence is important to identify whether the axes or the visual channel can be instantiated only once (e.g. bar chart x-axis) or multiple times (e.g. parallel coordinates x-axis). In fact, this model is technology-independent and used by the Visualization Wizard to generate the specific visualization code. We use in our Wizard the D3[6] visualization library and Google Charts[7] to create our visualizations but as mentioned above, it is possible to use other technologies.

Currently, the Visualization Wizard supports nine different charts and a table. For the integration of each new visualization, a generator needs to be implemented, which has well-defined interfaces and can be plugged-in to the Visualization Wizard easily.

Mapping Vocabularies: The mapping between both mentioned vocabularies, the RDF Data Cube and the VA Vocabulary, is a relation from dimensions and measures of the RDF Data Cube (i.e. cube components) to the corresponding axes and visual channels of the visualization. The mapping combinations will be found based on the structural compatibility and on the datatype compatibility between a RDF Data Cube and visualizations.

The number of the dimension and measures in a RDF Data Cube is unbounded. The possible combinations (i.e. in the format dimension: measure) for each RDF Data Cube are: (1) `1:1`, (2) `1:n`, (3) `n:1` and (4) `n:n`. The structural definition of a visualization represents, how many axes/visual channels the visualization has. To find a valid mapping, the VA Vocabulary has to suggest visualizations with the same structural definition like the structural definition of the corresponding RDF Data Cube. To clarify this, let us analyze the bar chart from the Figure 1: The bar chart has two axis, `x-axis` and `y-axis`, and can only visualize RDF Data Cubes with one dimension and one measure (`1:1`). The structural compatibility is not sufficient for a valid mapping, but also the datatype compatibility. The datatype compatibility is based on the primitive datatypes[8] (string, integer, float etc.) supported by the both vocabularies

[5]PAN: `http://pan.webis.de/`

[6]D3: `http://d3js.org/`

[7]Google Charts: `https://developers.google.com/chart/`

[8]Datatypes: `http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/`

(see Visualization Process). Since the RDF Data Cube may expose composite datatypes, these must be mapped to supported primitive datatypes.

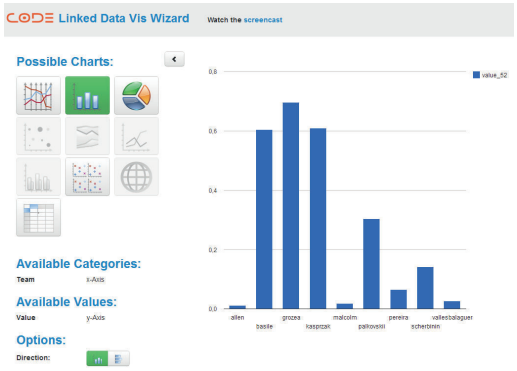

Fig. 1. The automatically generated visualization of PAN Data with *Team* as RDF Data Cube dimension and with a *Value* as RDF Data Cube measure.

Visualization Process: Based on the provided RDF Cube model, the Visualization Wizard proposes (1) visualizations and (2) possible variants of the mapping (see Fig. 2). The mapping is done by depicting dimensions and measures on the provided axes or on the visual channels of the visualizations. For instance, a bar chart consists of two axes: `x-axis` with a *string* and `y-axis` with a *decimal* datatype. Here, a dimension of datatype *string* will be mapped onto the `x-axis` and a measure of datatype *decimal* onto the `y-axis` (see Figure 1). However, if there are more dimensions or measures with the same datatype, we have various mapping variations for a visualization with axes which have the same datatype like these cube components. In this case (the option 2), the wizard creates a candidate table including all possible combinations between both models. The user can choose between different combinations, and for each combination, a specific visualization will be created and the provided data will be automatically visualized.

4 Conclusion and Future Work

The challenge of the first iteration in developing the CODE Visualization Wizard was to show that pitfalls of traditional visualization principles, such as the need for the manual work and high maintenance while visualizing datasets, can be effectively overcome by describing data and visualizations in dedicated vocabularies and by mapping these vocabularies. From the technical viewpoint, the main challenge was to automatically determine the right mapping between instances of the RDF Data Cube and the existing visualizations. Another, and

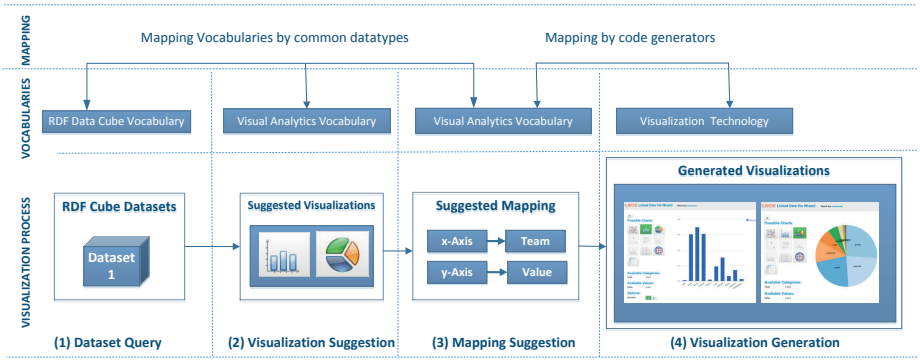

Fig. 2. Main parts of the Visualization Wizard: automated visualization process (bottom), vocabularies (middle) and mapping vocabularies (top). See live demo[3].

more serious challenge was to determine only valid suggestions among the provided visualizations.

The ongoing topics, which are parts of the project's next iterations, are (1) the investigation and the implementation of methods on how to use the previous user's knowledge (i.e. stored mappings) in order to effectively suggest mappings, (2) the extension of the automated visualization model for RDF Data Cubes with no explicit datatypes and (3) the implementation of refinement functionalities, like zooming, filtering etc.

The development of the prototype will continue throughout the rest of the year, leading to a final evaluation at the beginning of 2014.

Acknowledgement This work is being developed at the Know Center within the CODE project funded by the EU Seventh Framework Programme, grant agreement number 296150. The Know-Center is funded within the Austrian COMET Program-Competence Centers for Excellent Technologiesunder the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

References

1. Chan et al. Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration. IEEE Trans. Vis. Comput. Graphics, 14(6), 2008, 1213-1220.
2. Dadzie et al. Approaches to visualising linked data: A survey. Semant. web 2(2), 2011, 89-124.
3. Dumontier et al. Modeling and querying graphical representations of statistical data. Web Semant. 8(2-3), 2010, 241-254.
4. Seifert et al. Crowdsourcing Fact Extraction from Scientific Literature. Proc. of HCI-KDD 2013 Workshop, pp. 160-172, 2013.
5. Stegmaier et al. Unleashing Semantics of Research Data. 2nd Workshop on Big Data Benchmarking, 2012.