Visual Statistics Cockpits for Information Gathering in the Policy-Making Process

Dirk Burkhardt, Kawa Nazemi, Christian Stab, Martin Steiger,
Arjan Kuijper, and Jörn Kohlhammer

Fraunhofer Institute for Computer Graphics Research (IGD),
Fraunhoferstr. 5, 64283 Darmstadt, Germany
{dirk.burkhardt,kawa.nazemi,christian.stab,martin.steiger,
arjan.kuijper,joern.kohalhammer}@igd.fraunhofer.de

Abstract. A major step in ICT-driven policy making is information gathering. During this phase, analysts and experts have to deal with a high number of statistical data which they use as a basis to identify problems and find appropriate solutions. This paper introduces a statistical data model to support these analysts and experts. It allows for handling the complexity (i.e. the dimensions) of the data for the visualizations. In particular, it helps to use the same data for two-dimensional, but also multi-dimensional statistics visualizations. Based on this statistic data model we introduce an interactive approach of visual statistics cockpits. This results in highly interactive statistics visualization cockpits that enable both analysts and experts to improve problem assessment and solution finding.

1 Introduction

Nowadays, policy making is a process that is primarily performed offline, so that most of the already existing benefits of Information and Communication Technologies (ICT) are not yet considered. The biggest advantage today is that most countries are working on open-data portals to improve transparency of the government's work for the citizens. These open-data portals allow citizens to analyze the government's work through large list of indicators that are measured by public authorities. Therefore, interested individuals and organizations are developing more and more tools and sophisticated visualizations to allow a facilitated analysis, which allows citizens to make use of published data. Unfortunately, these open-data portals are not yet established in the policy making process of public authorities. Especially municipalities consider ICT and open-data today in a very limited way. Over the past years, the number of municipalities and politicians who understand the necessity to incorporate ICT in the policy making process has increased. They started to consider how ICT could be included to improve policy making process in their daily lives.

In order to develop ICT that help public authorities in the policy making, it is necessary to develop tools that help both analysts and experts to grasp the meaning of the available data. This can primarily supported through the use of adequate visualization

G. Bebis et al. (Eds.): ISVC 2013, Part II, LNCS 8034, pp. 86–97, 2013.

tools that allow for showing data in an adequate manner and enable users to get an understanding of a given problem. Furthermore, these tools need to be included in the existing policy making cycles of the public authorities. Considering adequate visualizations as well as inclusion into the given policy making processes allows for an effective policy creation. An approach that directly addresses visualization in the policy making process was published by Kohlhammer et al. [17]. They defined the most relevant visualization tasks for policy making: (1) information foraging, (2) policy design, and (3) impact analysis. They also addressed information foraging as the major task for visualization, because it includes both the problem identification and the solution finding.

Today, the general aspects of statistical data visualization are well known; there are also a couple of approaches using dashboards to allow for an improved overview of the data. A major challenge in the domain of visual analysis is the interactivity of statistics visualizations. In this paper we address this issue and introduce a novel design of a technical data model for statistical data, which is primarily designed for modern open-data portals. The main idea is to specify a data model that allows for usage in multiple types of statistics visualizations. In particular, it aims to bridge the gap of coupling two- to multi-dimensional statistics visualization with only one data model that holds all available data. Based on this data model, we outline an approach of visual statistics cockpits for interactive analysis. The main benefit of the technical statistics data model and the statistics cockpit approach is the support of analysts and domain experts in the policy making process. This is essential in the information gathering phase to identify problems and solutions just through analyzing the data.

2 Related Work

Interaction with statistical visualizations has been investigated since a couple of decades. It is one of the best researched topics in information visualizations sciences. The focus was mostly on how statistical data should be communicated in a graphical way, and how it will then be easier understood by the target users. For these target users a couple of simple but also expert visualization were developed.

The major challenge of using statistical visualizations in the domain of policy modeling is not the use of statistical visualizations in general, but it is the practical integration of appropriate visualizations with respect to the task, the data, and the user [18]. Especially the user and his level of expertise define whether the visualization is useful or rather confusing. Therefore, we give an overview of existing visualization approaches in the first part of this section and define for which user-type they will be adequate. In the second part of this section we introduce existing approaches of dashboards, which follow the idea of coupling multiple visualizations to provide a better overview. The disadvantage of single visualizations is that they are not appropriate for all analysis works and therefore different sets of visualizations can be beneficial for the user. In this section we introduce some examples of dashboards and describe their benefits for the analysis.

2.1 Statistics Visualization Approaches

Today, a large number of statistical visualizations exist, but most of them can be categorized into only four groups [4]. Visualizations offer their advantages in regard to the expected use-case. Consequently, there is no visualization that will be beneficial for all kinds of analysis work. Any visualization provides a facetted view on the data. Statistical data can be considered as lists or tables of nominal or ordinal data [4]. Additionally, the benefit of visualizations also depends on the user's behavior and his expertise in using visualization systems. An analyst can use visualizations which can show multidimensional data within complex visualizations, i.e. Pixel Matrix Displays [1] or Parallel Coordinates [2, 3]. In fact, this means that any visualization has to be selected in dependence on the use-case and the tasks that should be solved with it.

In the first category, the *Point-based methods* [4], visualizations make use of points, marks, or other symbolic signs for representing a data record. According to certain attributes, the visual representations of each record are placed on the screen to derive a visual representation of the whole data set. Depending on the selected attributes and the chosen layout method, point-based techniques are suitable to compare certain data characteristics, to identify outliers or irregularities in the data, to recognize relationships among data entities, and to identify unexpected or previously unknown clusters and patterns. Usually, each data record is projected from its n-dimensional space to a (lower) k-dimensional space (usually two- or three-dimensional) and the visual representation of the record is represented at the k-dimensional point on the screen. Example visualization representatives are (Parallel) Scatterplots [5] and (vectorized) RadViz [6]. Point-based visualizations, especially the more abstracted visualizations like the RadViz, are most suitable for advanced users. Next to the placement of the point in the coordinate system, also the shape, size, or color imply information and this metaphor is usually not easy to understand by non-advanced users.

Line-based techniques form the second category for visualizing statistical and multivariate data. They are often used to analyze financial or temporal data. Due to the high familiarity of the respective users, line-based techniques are an effective tool for common users to analyze and explore statistical data. In contrast to point-based methods that represent each data record as a symbolic representation, line-based visualizations are based on the idea of representing the values of a data record or dimension linked together in a straight or curved line. Thus, each line in a line-based visualization represents perceivable features of the given record. Consequently, these techniques are well-suited for identifying slopes, curvatures, and even crossings among multiple records. Typical representatives of line-based visualizations for statistical data and their different modes and goals are presented in the following subsections. The most established and best known representative for this kind of visualization are Line Charts. Line Charts are adequate for casual as well as advanced and expert users. They present two-dimensional data in an easily understandable form. However, other visualizations are rather designed for experts, like Parallel Coordinates [2,3] and Radial Visualization [7].

The third category, named *Region-based visualization techniques*, is used for multivariate data. They are often used to analyze financial data, to compare actual and target status, or to explore differences in experimental data. The basic idea of region-based visualization techniques is to represent data records as filled polygons or regions on the screen. Usually instantiations of region-based techniques incorporate different properties of the given data into the visual design of the polygons to convey additional values and data characteristics to offer the possibility of comparing different features of the data. For instance, the size, the shape, or the color of the visual representation of a data record can be utilized for visually representing additional dimensions of the data set. Due to the ability of the human perception – which enables an effective differentiation of the length or the size of presented polygons – region-based visualizations are successfully applied for representing and analyzing different quantitative information encoded in the data. Typical representatives of region-based visualizations include bar charts, tabular displays like heat-maps [8], and table lenses [9]. These kinds of visualization are primary designed for advanced users. Implementations that can be used by casual users or – for data with a higher level of complexity – by expert users, exist as well.

The forth category is the *Combination of Techniques*. Point-based, line-based, as well as region based techniques utilize a specific graphical metaphor to visualize statistical or multivariate data sets. Hybrid visualization techniques incorporate several techniques of the methods presented in previous sections for obtaining a meaningful representation of the given data. Typical representatives of hybrid visualization techniques for multivariate data are Dense Pixel Displays[10] and Theme River [11]. Because of the combination of different approaches, most of the visualizations become so complex that the target users are usually advanced and expert users.

2.2 Statistics Dashboards

The use of dashboards is a common approach to reduce the problem of insufficient visualization for a certain task. Because of the combination of different visualizations, it can be ensured that at least one adequate visualization is shown to the user. Furthermore, the use of multiple visualizations allows showing the data from different perspectives, and thus providing an adequate overview.

The policy making process, especially the analysis phases, counts as a very heterogeneous task. Depending on the goal of the analysis the number of data elements that has to be analyzed varies. This comes along with strong deviations in the data dimensionality, and the resulting complexity in the visualizations. The more indicators need be conducted, the more complex the visualizations become. The use of dashboards can help to reduce the complexity, as next to detailed visualization also a very abstract overview visualization can be provided. This allows a top-down analysis (similar to Shneiderman's information seeking mantra [12]) and acts as a low-barrier entrance into the data analysis.

In the policy modeling domain two well-know dashboard approaches exist that make use of a combined set of visualizations. The *OECD eXplorer*[1] [13, 19] allows for analyzing various statistical data from OECD countries. It therefore provides a set of visualizations to show indicators, such as the GDP, population development, and education. For this purpose, the OECD Explorer uses next to traditional chart visualization also some modern visualization techniques like scatter plots, parallel coordinates, and colored geographical visualizations. The dashboard consists of an orchestration of visualizations. Only the statistical visualization can be replaced with a single other visualization. Thus, the OECD Explorer provides a first interactive approach to analyze statistical data, but it is very restricted in the number and the type of available visualizations.

The second dashboard is the *EZB Inflation Dashboard*[2], which allows for an analysis of inflations of different European countries. It provides a dashboard consisting of three common chart visualizations and a map visualization of Europe. With this dashboard the user can analyze and compare the inflations in the European countries in an interactive manner. To make detailed analysis, any visualization can be switched to full screen.

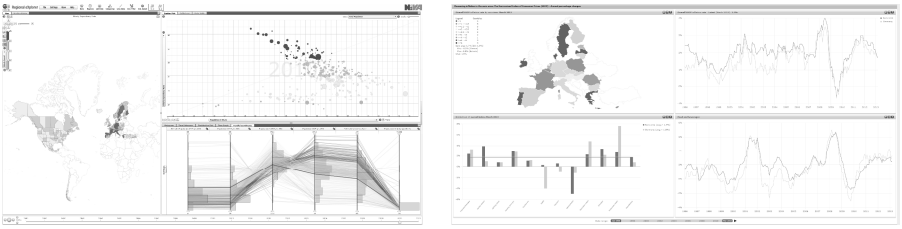

Fig. 1. On the left a screenshot of the OECD eXplorer is shown with its different visualizations. On the right a screenshot of the EZB Inflation Dashboard is shown, with its four fix-defined visualizations orchestration.

To provide interactivity it is necessary to provide the user visualizations that allow for an in-depth analysis, but also to allow interactivity through a dashboard metaphor. Traditional chart visualizations are well-known and use simple metaphors, but they are limited in single-use scenarios regarding interactivity. Modern visualizations, like parallel coordinates, are more complex, but allow for an improved in-depth analysis. Consequently, it seems beneficial to provide an analysis visualization system that contains a mixed set of traditional chart visualization, as well as modern analysis visualizations. To enhance this general approach, it also seems to be beneficial, when the visualizations are presented with a dashbod metaphor. Here the user can choose the appropriate visualization for his work. A major benefit of a dashboard with coupled visualizations is the higher interactivity. This metaphor should be used and

[1] Available on: `http://stats.oecd.org/OECDregionalstatistics/` (accessed on 29/04/2013)

[2] Available on: `http://www.ecb.europa.eu/stats/prices/hicp/html/inflation.en.html` (accessed on 29/04/2013)

extended through a higher flexibility. The described approaches use static orchestrated visualizations. From the interactivity perspective it seems beneficial, when the cockpit can be completely orchestrated by the user depending on his behavior and work task. Therefore however, a concept is required that allows for the unification of the existing two- to multidimensional data in a single technical data model. If all kinds of visualization are able to handle the same technical data model with the statistic information, a flexible cockpit can be designed.

3 Concept for Visual Statistics Cockpits

The design of a statistics visualization dashboard that should include various types of static visualizations, consisting of two- and multidimensional visualizations needs to introduce a concept to handle the data. Hereby we have to deal with the existing open-data portals and how this data can be used in these two- to multidimensional visualizations.

Existing open-data portals, e.g. EuroStat, GOVDATA, Data.gov and Data.gov.uk, mostly structure the data into a hierarchy of topics. For each topic a couple of so-called indicators are aligned. Furthermore, each indicator is defined by:

- The name of the indicator, e.g. GDP, public growth, or public density.
- An assignment to a geographical region, i.e. a country, state/province, municipality, or city.
- A time-based data table, which consist of the indicator value by the measured time.
- Optional additional meta-information about the indicator, such as a description of influencing indicators or the used unit.

Based on these given structures, a technical data model that is adequate for the different two- and multidimensional statistics visualizations needs to be designed. Additionally, the data model needs to support interactive approaches, such as the option to link visualizations and to allow a further exploration through the data.

3.1 Concept for a Statistic Data Model

The technical data model is a basis that allows an interactive analysis. Therefore, it must be ensured that it provides the ability to navigate through the data with all kinds of visual statistic representations.

Based on the already mentioned structure of the existing open-data platform, we identified the relevant parameters that can be changed in the visualization process to explore the data. A change of the parameters is required when the user wants to select parts of the entire data set or when the visualization is limited, e.g. on just two dimensions (see also Fig. 2).

The major grouping option is defined by the indicator name, the geographic location, and time. These properties are basic elements that every indicator comprises.

It allows for using 2-dimensional visualization on multidimensional data, just by reduction through one of the grouping options.

Additionally, an advanced grouping option is also considered. This procedure is only available on complex data sources which provide such additional meta-information. An example is EuroStat over the SDMX API. SDMX[3] is one of the most established sharing formats for statistical data and meta-information. The structure of meta-information is similar to Linked-Open-Data [14] and can be used to provide an enhanced navigation through the statistical data. In this paper we just consider the major grouping options, because most public data sources are currently in the progress to include additional meta-information. Furthermore, if they have such meta-data, they often use very heterogeneous structures which make it difficult to combine them from multiple data sources.

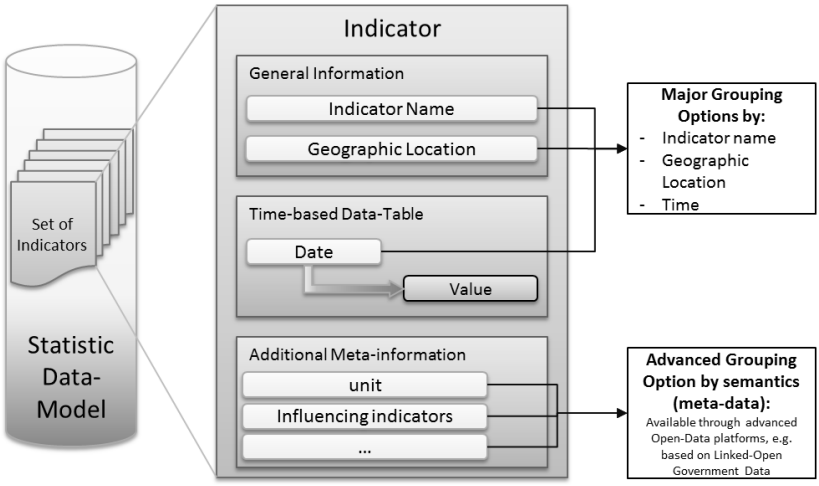

Fig. 2. The structure of the statistic data model is shown, including the main properties of an indicator which is primarily used for the grouping process. An advanced grouping can also be done through additional meta-information

3.2 Extension and Reduction of Data-Complexity for the Statistical Visualizations

A general challenge in coupling various visualizations is the differently supported dimension of the underlying data. A multi-dimensional visualization can show all available dimensions, but simple visualizations, like pie charts, support only two dimensions. As a consequence, the application on a single data base or data model then proves to be difficult in general.

In the previous section we introduced a novel design of a technical data model for statistical data. Based on this concept we show how the high-dimensional data, that

[3] More information about the SDMX format on: `http://sdmx.org` (accessed on 29/04/2013).

are available in the presented data model, can be abstracted so that also a low-dimensional visualization can show the same data. This is achieved through a simple limitation of the dimensionality. In this section we focus on the major grouping options.

In the statistical data model we mentioned the indicator's name, the geographic location, and the time data table as major grouping options. We can now adjust the dimensionality through enabling, disabling, or selection/filtering of entries based on the data model. For multi-dimensional data all available indicators with their geographic location and time data table can be shown. In contrast, two-dimensional visualizations reduce the data on just one concrete indicator (e.g. GDP) and for a specific date (e.g. 2012). The final visualization then shows the GDP from 2012 for all available locations. As a consequence, any developed visualization can reduce the complexity in dependence of the supported dimensions. By default, the visualization reduces the complexity in its own way, but it is also possible that the user defines what information needs to be included. This means that the visualization generally can be configured during the interaction process. This also includes changes on the presentation, so that e.g. a line chart can be changed into an area or plot chart. It is also possible that visualizations can be decoupled to allow for a comprehensive view on the data.

Based on this simple data model, different kinds of statistical visualizations can be used on the same data model and therefore on the same data source.

3.3 Creations of Visualization Cockpits

To allow for an effective analysis, the personal orchestration of different visualizations is a beneficial approach. We entitle this visualization orchestration ability "cockpit", because it supports to control the view on the data. In general, the terms "cockpit" and "dashboard" [15, 16] are synonymously used, but we prefer the term cockpit which focuses more on an active use in a very complex environment. For this purpose, the cockpit provides a higher degree of interaction and opportunity to orchestrate a personalized cockpit so that a given task can be solved more efficient.

As a general design we consider a list of data sources, i.e. EuroStat and some local municipality data bases in the top. The user can switch between these data sources. Next to the data base options, the user can also search for one or more indicators that should be visualized. On the right side, a set of visualizations (simple chart visualization, as well as complex visualizations like Parallel Coordinates) is displayed. In the center, the user can drag and drop visualizations to analyze the data. Any visualization can be configured in detail, e.g. decoupling the visualization for a comparative view. In order to focus on the most relevant visualization, the user can also resize the visualizations as needed.

4 Implementation

In a first prototype of a statistics cockpit we implemented some basic statistical visualization as examples, e.g. line chart, pie chart, bar chart, and a data table. On the top, the user can search an indicator. On the right side the user can drag and drop the preferred statistical visualization on the cockpit.

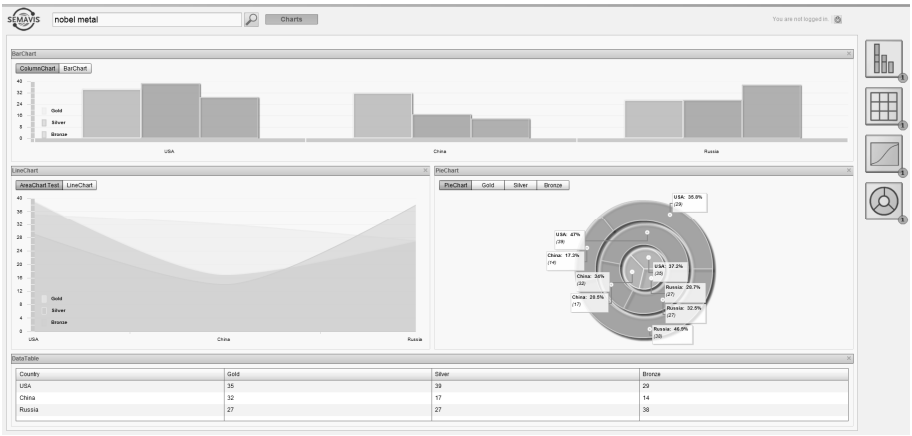

Fig. 3. The picture shows the implemented visual statistics cockpit of noble resources data. The cockpit can be individually composed and resized.

The prototype is currently not yet fully implemented, but it can already be shown that it provides the ability to analyze the data in a sufficient way. Especially in the information foraging phase [17] where analysts aims to identify problems and solutions, a multi-angle view on the data supports the work.

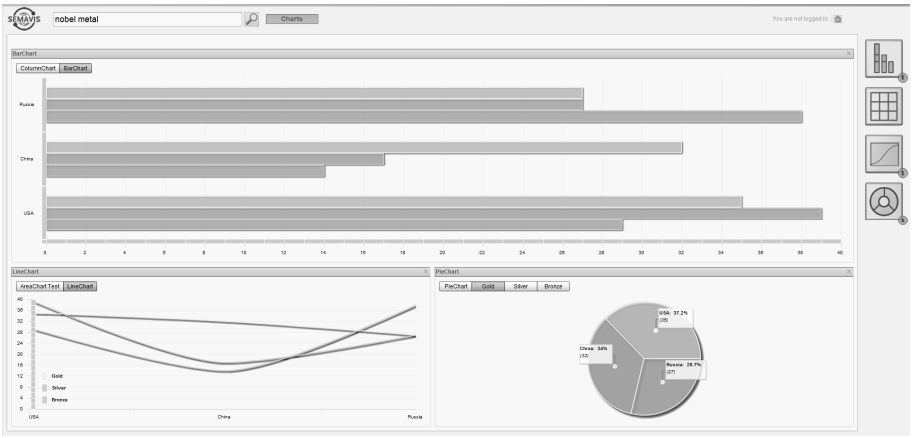

Fig. 4. The picture shows another cockpit view of noble resources data. As can be seen here, the cockpit can be customized and resized easily to the user's preferences and requirements.

We propose to include this statistical analysis in process-driven environment, to support the user to solve his task more efficient [20, 21]. This enables the user to get visualization or other kinds of visual tools in dependence on the current task in the process.

5 Discussion

With this prototype we have built the (conceptual) foundation for an interactive statistics cockpit. Through the use of established and modern analytical visualizations, the user is able to analyze the data. An important question is in how far interactivity can be supported in an elementary way. The goal for explorative analysis depends on the interactivity and the user experience. It is currently difficult to achieve an explorative character in the analysis, because the users have a clear goal they want to achieve. An approach for explorative analysis needs increasing interactivity of a system and more space for just "playing" around with the data, but currently no established strategies are available.

Another open question is the coupling of SDMX meta-information from different data sources. The SDMX meta-information are similar to LOD data source (especially *DBpedia*[4]), but also for LOD no full working and optimal mapping approach does exist. A contribution of such a coupling of SDMX meta-information and other LOD information (e.g. dbpedia) is the ability to show additional explanations to users on demand. So, the user gets, among other, further information about the used unit or how the data items are imposed. So the coupling can act as a possibility to enrich the information about the available statistical data.

6 Conclusion

This paper introduces a concept to provide interactivity in statistical visualizations, which are orchestrated in a statistics cockpit. For this purpose, a new technical statistical data model was introduced that allows for extending or reducing the dimension of the data for different existing statistics visualizations. This approach allows for using the same data on different visualizations types and provides an improved linking ability. The linking is beneficial to support the provision of statistics cockpits and to give a well-organized overview on the data. Such orchestrated cockpits showing the same data in different forms create a multi-angle view on the data.

This overview on the data is useful in the process of policy modeling. Both analyst and expert have to deal with a large number of data in the process of information gathering in order to identify problems and to find possible solutions. The ability to interact with the data allows for an improved solution finding. Additionally, the orchestration of various visualizations also allows for an improved insight in the data. Thus, a problem can be better analyzed and thus better understood as it would be possible in a single visualization.

In future work we aim at bringing the SDMX meta-information with LOD together. The primary goal is to increase the interactivity component. We plan to include better explanation of the statistical data e.g. about the usage of some indicators. For this, some premature works exist, e.g. the linking approach described in [14], which can be enhanced especially in the visual integration.

[4] DBpedia is a knowledge data-base in Linked-Open Data (LOD) structure, and it is available under: `http://dbpedia.org` (accessed on 29/04/2013)

Acknowledgements. This work has been carried in the FUPOL project, partially funded by the European Commission under the grant agreement no. 287119 of the 7th Framework Programme. This work is part of the SemaVis visualization framework, developed by the Fraunhofer IGD (http://www.semavis.com). SemaVis provides a comprehensive and modular approach for visualizing heterogeneous data for various users.

References

1. Hao, M., Dayal, U., Keim, D., Schreck, T.: A Visual Analysis of Multi-Attribute Data Using Pixel Matrix Displays. In: Proc. VDA 2007 (2007)
2. Yuan, X., Guo, P., Xiao, H., Zhou, H., Qu, H.: Scattering Points in Parallel Coordinates. IEEE Transactions on Visualization and Computer Graphics 15(6), 1001–1008 (2009)
3. Heinrich, J., Weiskopf, D.: Continuous Parallel Coordinates. IEEE Transactions on Visualization and Computer Graphics 15(6), 1531–1538 (2009)
4. Ward, M.O., Grinstein, G., Keim, D.: Interactive Data Visualizations: Foundations, Techniques, and Applications. Taylor & Francis Ltd. (2010)
5. Viau, C., McGuffin, M.J., Chiricota, Y., Jurisica, I.: The FlowVizMenu and Parallel Scatterplot Matrix: Hybrid Multidimensional Visualizations for Network Exploration. IEEE Transactions on Visualization and Computer Graphics 16(6), 1100–1108 (2010)
6. Sharko, J., Grinstein, G., Marx, K.A.: Vectorized Radviz and Its Application to Multiple Cluster Datasets. IEEE Transactions on Visualization and Computer Graphics 14(6), 1077–1427 (2008)
7. Draper, G., Livnat, Y., Riesenfeld, R.F.: A Survey of Radial Methods for Information Visualization. IEEE Transactions on Visualization and Computer Graphics 15(5), 759–776 (2009)
8. Wilkinson, L., Friendly, M.: The History of the Cluster Heat Map. The American Statistician 63(2), 179–184 (2009)
9. Rao, R., Card, S.: The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence, CHI 1994, pp. 318–322 (1994)
10. Keim, D.A., Kriegel, H.-P., Ankerst, M.: Recursive pattern: a technique for visualizing very large amounts of data. In: Proceedings of the 6th Conference on Visualization 1995, pp. 279–286 (1995)
11. Havre, S., Hetzler, E., Whitney, P., Nowell, L.: ThemeRiver: visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics 8(1), 9–20 (2002)
12. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings IEEE Symposium on Visual Languages, pp. 336–343 (1996)
13. Jern, M.: Collaborative web-enabled geoanalytics applied to OECD regional data. In: Luo, Y. (ed.) CDVE 2009. LNCS, vol. 5738, pp. 32–43. Springer, Heidelberg (2009)
14. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic Statistics: Bringing Together SDMX and SCOVO. In: Proceedings of LDOW, Raleigh, North Carolina, USA (2010)
15. Few, S.: Information Dashboard Design. Overview article about Dashboards (2012), `http://blogs.ischool.berkeley.edu/i247s12/files/2012/01/Dashboard-Design-Overview-Presentation.pdf`

16. Duval, E.: Attention please! learning analytics for visualization and recommendation. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK 2011, pp. 9–17. ACM, New York (2011)
17. Kohlhammer, J., Nazemi, K., Ruppert, T., Burkhardt, D.: Toward Visualization in Policy Modeling. IEEE Computer Graphics and Applications 32(5), 84–89 (2012)
18. Burkhardt, D., Nazemi, K., Sonntagbauer, P., Sonntagbauer, S., Kohlhammer, J.: Interactive Visualizations in the Process of Policy Modeling. In: Proceedings of IFIP eGov 2013. GI-LNI (2013)
19. Jern, M.: (OECD), What does OECD eXplorer enable you to do? An introduction to its main features. In: Handbook of the OECD eXplorer (2009), http://www.oecd.org/gov/43142629.pdf
20. Burkhardt, D., Ruppert, T., Nazemi, K.: Towards process-oriented Information Visualization for supporting users. In: Proceedings of 15th International Conference on Interactive Collaborative Learning, ICL 2012, pp. 1–8 (2012)
21. Burkhardt, D., Nazemi, K.: Dynamic process support based on users' behavior. In: Proceedings of 15th International Conference on Interactive Collaborative Learning, ICL 2012, pp. 1–6 (2012)