Chapter 14
Narrative Visualization of Open Data

Philipp Ackermann and Kurt Stockinger

Abstract Several governments around the globe have recently released significant amounts of open data to the public. The main motivation is that citizens or companies use these datasets and develop new data products and applications by either enriching their existing data stores or by smartly combining datasets from various open data portals.

In this chapter, we first describe the development of open data over the last few years and briefly introduce the open data portals of the USA, the EU, and Switzerland. Next we will explain various methods for information visualization. Finally, we describe how we combined methods from open data and information visualization. In particular, we show how we developed visualization applications on top of the Swiss open data portal that enable web-based, interactive information visualization as well as a novel paradigm—narrative visualization.

1 Introduction to Open Data

The idea of freely sharing open data has been around for several decades. For instance, the World Data Center[1] developed a concept for open access to scientific data in 1957–58. However, the open data movement for access to public data has only recently gained worldwide traction (Bauer and Kaltenböck 2011). One of the main drivers of the movement[2] is Tim Berners-Lee, who is often considered as the father of the World Wide Web. The main goals are to make local, regional, and national data electronically available to the public and to lay the foundations for different actors to build rich software applications upon them.

[1] www.icsu-wds.org

[2] Open Data Handbook: www.opendatahandbook.org

P. Ackermann · K. Stockinger (✉)
ZHAW Zurich University of Applied Sciences, Winterthur, Switzerland
e-mail: stog@zhaw.ch

© Springer Nature Switzerland AG 2019

M. Braschler et al. (eds.), *Applied Data Science*,
https://doi.org/10.1007/978-3-030-11821-1_14

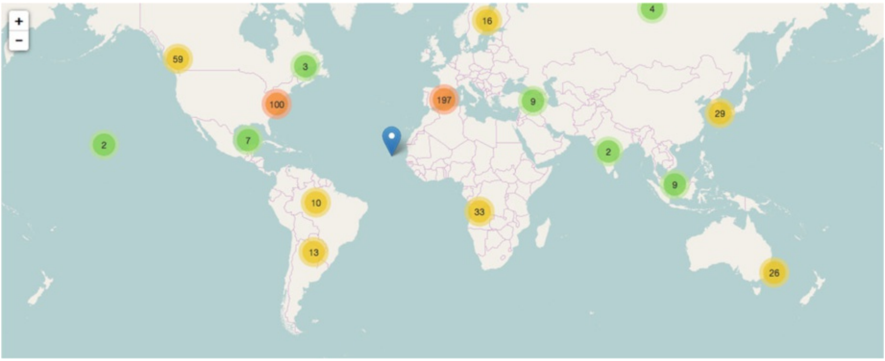


Fig. 14.1 List of data catalogs curated by experts around the world (http://dataportals.org/)

Table 14.1 Open data portals of the USA, the EU, and Switzerland as of April 2017

Open data portal	Provider	Number of datasets	Number of applications
data.gov	US Government	1,92,322	76
	European Union	10,702	70
https://opendata.swiss	Swiss Government	2169	30

Another important driver in the open government data movement is “The Memorandum on Transparency and Open Government” signed by US President Barack Obama shortly after his inauguration in January 2009 (Orszag 2009). The aim was to establish a modern cooperation among politicians, public administration, industry, and private citizens by enabling more transparency, democracy, participation, and collaboration. In European countries, Open Government is often viewed as a natural companion to e-government (Bauer and Kaltenböck 2011).

Figure 14.1 shows an aggregated number of open data catalogs curated by experts around the world. We can see that the major activities are in Europe and the East Coast of the USA.

Table 14.1 shows some facts about the open data portals provided by the USA, the European Union, and Switzerland. We can see that the US portal contains 1, 92, 322 datasets, while the portals of the European Union and Switzerland contain 10,702 and 2169, respectively. More interesting from our viewpoint, however, is the utilization of these datasets and thus the applications that are built upon them. The web portals currently list 76, 70, and 30 applications for the USA, the EU, and Switzerland, respectively. The applications are very diverse and are in the areas of health, finance, environment, government, etc. These exemplary applications demonstrate the great potential in harvesting open data and thus generating either new business models or new services for citizens.

Different countries pursue different strategies with Open Government Data (Huijboom and Van den Broek 2011). Whereas the emphasis of the USA is on transparency to increase public engagement, Denmark, for example, underscores the

Table 14.2 Top 10 ranking open government data by country as of 2017 (http://index.okfn.org/)

Rank	Country	Score (%)
1	Taiwan	90
2	Australia	79
2	Great Britain	79
4	France	70
5	Finland	69
5	Canada	69
5	Norway	69
8	New Zealand	68
8	Brazil	68
10	Northern Ireland	67

opportunities that open data offers for the development of new products and services. The UK explicitly mentions the use of open data to strengthen law enforcement.

The Global Open Data Index[3] annually measures the state of open government data around the world. The index measures the level of conversion to open data based on datasets provided by different areas such as national statistics, government and budget, etc. (see Table 14.2). According to the 2015 ranking, Taiwan is leading ahead of the UK and Denmark.

Open data is stored in portals under various formats such as comma separated values (CSV), PDF, or text files. However, the main storage technologies and APIs are based on two major frameworks:

- RDF (Resource Description Framework) and its query language SPARQL
- CKAN (Comprehensive Knowledge Archive Network)

RDF and SPARQL are the main technologies used for the semantic web as well as the Linked Data[4] movement. In RDF, every data item is stored as a triple of subject, predicate, and object, and enables linking objects on the web via URIs (uniform resource identifiers).

SPARQL is the query language for accessing data stored in RDF. It can be considered as the equivalent of SQL for relational databases.

RDF and SPARQL are used as the underlying technology for the open data portals of the European Union.

CKAN is an open source data catalog for storing and distributing open data developed by the Open Knowledge Foundation.[5] CKAN is used by the open data portals of the USA, the UK, and Switzerland.

In principle, the above-mentioned storage catalogs are not compatible. However, there exist SPARQL extensions for CKAN, which enable querying data stored in CKAN via SPARQL. An advantage of RDF/SPARQL over CKAN is that it is used by a much larger community, in particular by Linked Data, as discussed previously.

[3] http://index.okfn.org/

[4] http://linkeddata.org/

[5] https://okfn.org/

Moreover, CKAN is merely a storage archive solution, while RDF/SPARQL provides standards for storing and querying data embedded in a rich semantic framework.

2 Visualization Techniques

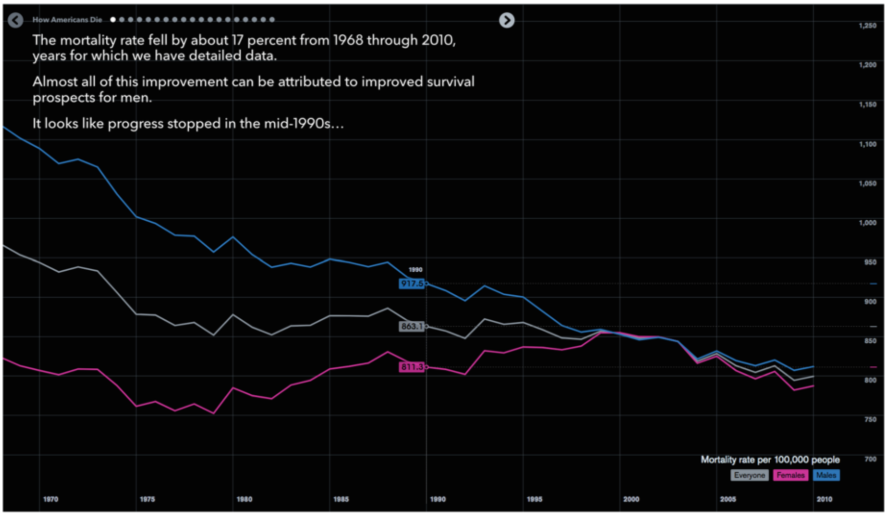
The primary goal of visual presentations of data is to enable the discovery and mediation of insights. *Data visualization* supports users in intuitively exploring the content of data, identifying interesting patterns, and fosters sense-making interpretations. Starting with a table of numbers is not an efficient way of interpreting data—an experience that is also commonly expressed in the idiom “a picture is worth a thousand words.” Complex information can only be slowly digested from rows of tabular data. By using graphical representations, we leverage the fact that our visual perception and our human cognition processes are more effective (Ware 2012).

In data-driven projects, analysts often map table-based datasets to *data graphics*. Embedded data visualization tools in MS Excel or R provide standard chart types such as bar charts and scatter plots to easily generate data graphics. Visual discovery with data graphics takes advantage of the human’s capacity to grasp a lot of information and identify patterns from visual representations of abstract data. Data graphics is already helpful with small amounts of data and becomes essential when datasets get bigger. Visually investigating data supports therefore discoveries driven by observation during exploratory analysis.

If the explored data sources reflect a certain complexity, analysts would like to interactively browse through the underlying databases. This demands for interactive *information visualization* (Shneiderman 1996) that enables users to ask additional questions and to dive into the underlying data within an elaborated information architecture. Because database schemas are not per se capable to evoke a comprehensible mental image, an information architecture transforms data sources toward mental models. We will provide a concrete example in Sect. 4.

After several iterations, the analyst may discover some findings worth presenting. Additional efforts are needed to make insights available to a wider audience. By providing an elaborated content structure within a semantic vocabulary and specific interaction patterns to filter, drill-down, and navigate through the information space, users better understand the presented content within an interactive information visualization.

If the underlying data sources provide the basis for several insights, analysts would like to communicate findings in a guided manner while maintaining the interactive exploration to users. Documenting the discovery of insights by guiding users through data landscapes is supported by *narrative visualization* techniques (Segel and Heer 2010). By presenting curated data exploration, users can step through a story, but also interactively explore the underlying datasets to gain their own insights. Data-driven journalism applies narrative visualization to present

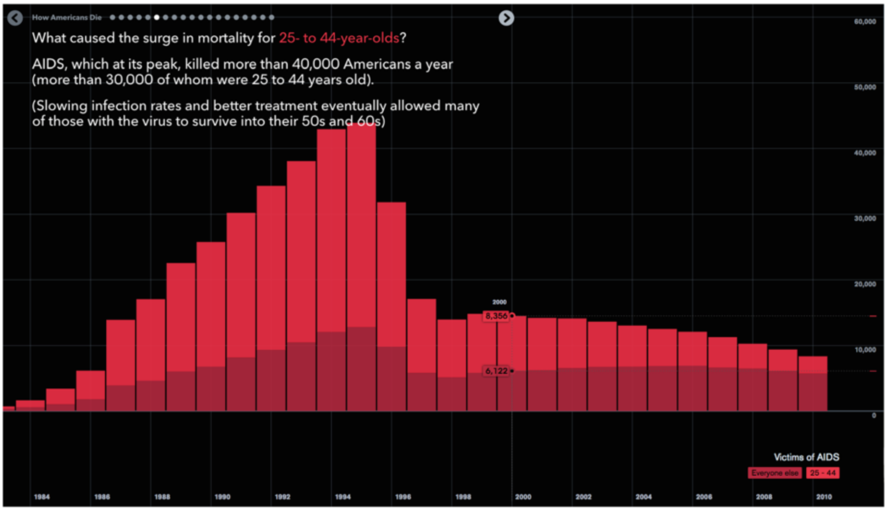


The mortality rate fell by about 17 percent from 1968 through 2010, years for which we have detailed data.

Almost all of this improvement can be attributed to improved survival prospects for men.

It looks like progress stopped in the mid-1990s…

Fig. 14.2 Excerpts from a sequence in a narrative visualization (www.bloomberg.com/dataview/2014-04-17/how-americans-die.html). The figure shows the declining mortality from 1968 to 2010. For men the improvement was most significant



What caused the surge in mortality for 25- to 44-year-olds?

AIDS, which at its peak, killed more than 40,000 Americans a year (more than 30,000 of whom were 25 to 44 years old).

(Slowing infection rates and better treatment eventually allowed many of those with the virus to survive into their 50s and 60s)

Fig. 14.3 Excerpts from a sequence in a narrative visualization (continued). The figure attributes the increased mortality rates in the 1980s and 1990s for 25–44-year-olds to AIDS

interactive visual stories. A good, illustrative example that was produced by Bloomberg is shown in Figs. 14.2 and 14.3.

Narrative visualization depicts key insights using storytelling with animations and therefore enhances data graphics and information visualization. Applications of

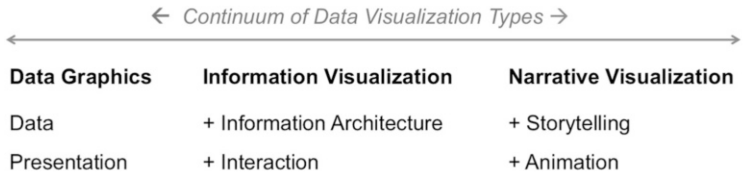

← Continuum of Data Visualization Types →
Data Graphics
Information Visualization
Narrative Visualization
Data
+ Information Architecture
+ Storytelling
Presentation
+ Interaction
+ Animation

Fig. 14.4 Types of data visualizations

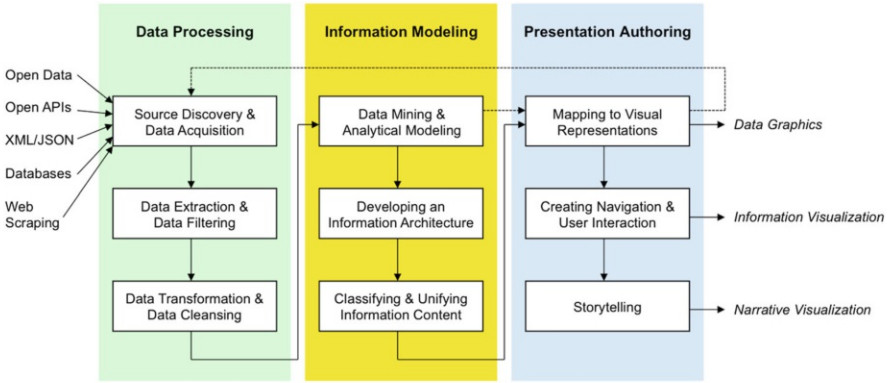

Data Processing
Information Modeling
Presentation Authoring
Open Data
Open APIs
XML/JSON
Databases
Web Scraping
Source Discovery & Data Acquisition
Data Extraction & Data Filtering
Data Transformation & Data Cleansing
Data Mining & Analytical Modeling
Developing an Information Architecture
Classifying & Unifying Information Content
Mapping to Visual Representations
Creating Navigation & User Interaction
Storytelling
Data Graphics
Information Visualization
Narrative Visualization

Fig. 14.5 Data visualization workflow

data visualization often use a mix of data graphics, interactive information visualization, and narrative visualization (Fig. 14.4). Applications using interactive and narrative visualizations need custom development of specific software to present findings to end users. To reach many users, such visual presentations are often deployed via the Internet or via an organization-specific intranet. Development libraries for web-based interactive data visualization such as Shiny[6] (Beeley 2016) and d3.js[7] (Bostock et al. 2011) provide core visualization technologies and many examples in the continuum of the three data visualization types.

3 Data Visualization Workflow

The workflow for developing data visualizations comprises several steps for data processing, information modeling, and presentation authoring (see Fig. 14.5). Although the workflow process is mainly sequential, typically many iterations for refining the deliverables are needed to create convincing data visualization results.

[6] shiny.rstudio.com

[7] www.d3js.org

Data Processing Phase A first step in data processing includes the discovery, identification, and acquisition of available data sources. With a certain interest and goal in mind, analysts make decisions about extracting and filtering relevant datasets. By combining different data sources their diverse content may be transformed to equal number formats and unit measures or may be converted by resampling to common time intervals. During the transformation and cleansing phase, a strategy needs to be developed to deal with wrong and missing values.

Information Modeling Phase Without knowing the content of data sources upfront, analysts explore and analyze the content with an open mindset oblivious to what exactly they are searching for. Exploring unknown datasets means interactively performing a sequence of queries to gain insight and understanding. During such data mining iterations an analytical model is elaborated by calculating statistical results that enhance the raw data. Additionally to the analytical model focused on numerical calculations, textual data needs to be brought into an information architecture (Rosenfeld and Morville 2015) by applying a unified content structure and using an explicit, typically domain-specific ontology that standardizes on used vocabulary and classifications.

Presentation Authoring Phase Once a unified information model is established, decisions have to be made how to map data to visual representations. The generation of data graphics compromises visual mappings to shapes, positions, dimensions, rotations, colors as well as the overall chart layout. If interactive information visualization should be provided, user interaction for navigation, filtering, and drill-downs need to be developed to communicate complex insights. The visual information seeking mantra by Ben Shneiderman “Overview first, zoom and filter, then details on demand” (Shneiderman 1996) requires some preprocessing in order to structure and aggregate the data to create overviews and drill-down navigation paths. If narrative visualization is aimed to additionally achieve storytelling, animations become part of the data visualization workflow.

4 Visualization for Exploring Open Data

By working through the data visualization workflow, data visualization itself is used to iteratively improve the results. Visual representations of data are helpful in gaining insights and in making evidence-based decisions to reach a convincing information presentation. Data exploration and analysis tools such as Tableau[8] and QlikView[9] provide interactive environments to inexpensively run visual analytics on a variety of data sources.

[8] tableau.com

[9] qlik.com

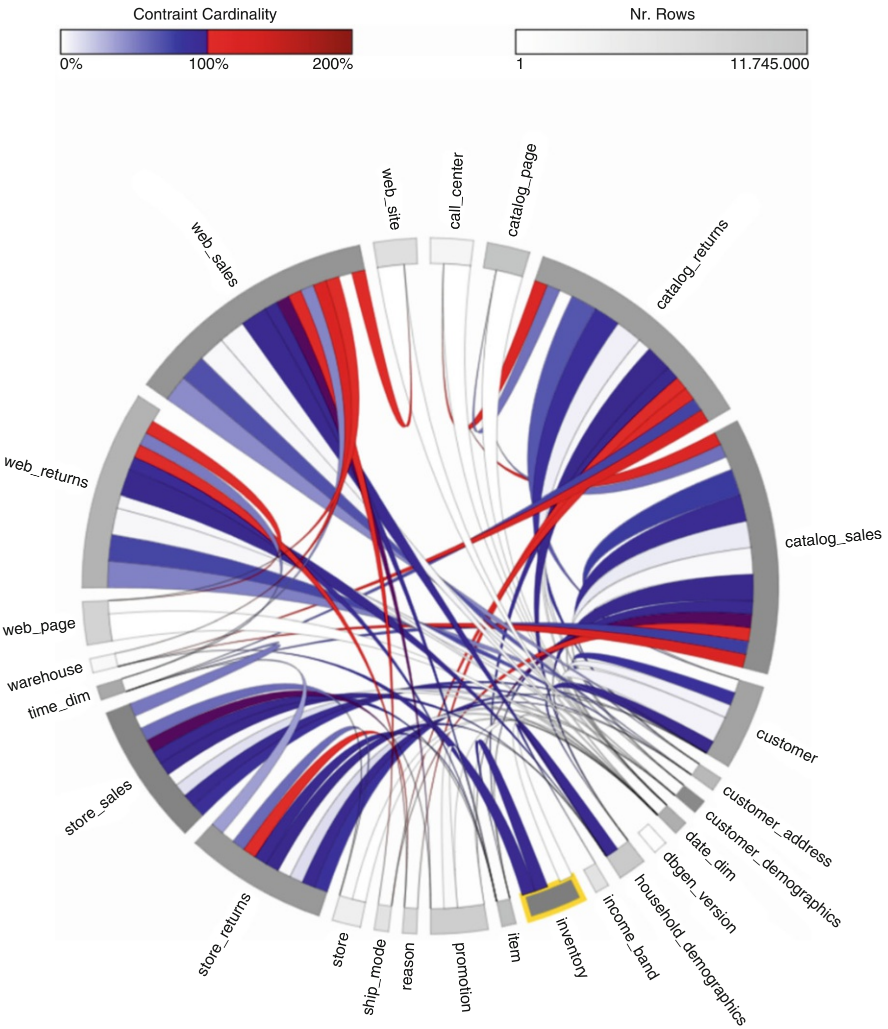


Fig. 14.6 Visual exploration of data structure and data content (Waser 2016). © 2016 Zurich University of Applied Sciences, reprinted with permission

The content of open data sources is often represented as relational tables in a database. In order to get familiar with an unknown database, Waser (2016) developed a visualization tool for SQL databases (see Fig. 14.6) to visually explore the structure and the content in parallel. The tables of the database are visually arranged in a circular layout showing the relations between tables that are color-encoded as arcs. Selecting tables and/or relations in the visual diagram automatically generates queries with corresponding joins. These visual queries are executed on the fly and the query results are shown as tables that can be sorted and filtered.

Providing such easy to use visual queries, users can investigate the data and gain insights in the structure and content of the database. On demand, histograms provide statistical evaluation of the content of table fields. Such visual data mining and exploratory data analysis tools help data analysts to make discoveries of patterns and trends driven by observations. These insights will then build the basis for interesting features worth presenting to a wider audience by developing specific information visualization applications.

5 Narrative Visualization for Presenting Open Data

We now describe how we used the Swiss open data portal to present the situation of health care services in Switzerland. Before the data was visualized, several data preprocessing steps were required.

In our case, we used the CKAN API provided by the Swiss open data platform to access the data. This API gives access to the metadata about the datasets, such as data owner, time of upload, description about the datasets, etc. However, it turns out that in most cases the metadata could not be directly used for visualization since the attribute description or units were missing. In this case, we needed to add the metadata manually.

Most of the data provided at the Swiss open data platform are in ODS-format, which cannot be directly read by the visualization framework D3.js. Hence, we needed to transform all datasets into tab-separated values (TSV) to enable visualization. Note that each additional data transformation step might introduce potential data quality problems that need to be handled explicitly by the application developers.

The next step was to integrate various datasets that contain information about heart attack, high blood pressure, body mass index (BMI), as well as public health data. An example of the entity relationship diagram is given in Fig. 14.7 (Montana and Zahner 2015). The main challenge here was that the datasets and the time ranges are of different granularity. Hence, we needed to standardize and harmonize the data before we could link it with other datasets. These are typical data warehousing tasks that all companies need to perform when they integrate data from various, heterogeneous datasets.

Once these data preprocessing steps were finished, we proceeded to tackle the actual task, namely, building interesting visualization stories. In particular, we were interested in the occurrence of high blood pressure, heart diseases, and overweightness of people in various age groups. We also wanted to know the differences in females and males. The goal was to build a sequence of interactive visualizations with intuitive narratives that guide a user from one visualization page to the next one. This approach gives the users a similar experience to reading a book or an article. The added value of interactive narrative visualization is that users can follow different lines of thought in a story by interacting with the visualized content.

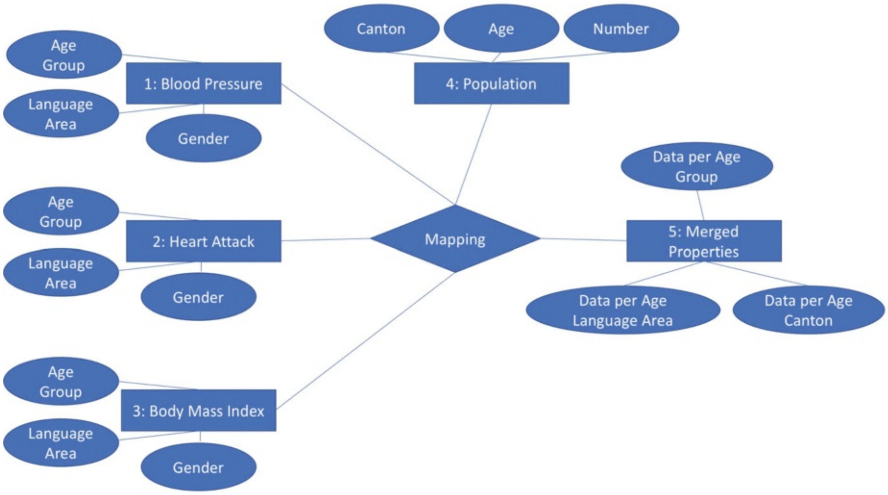


Fig. 14.7 Entity Relationship Diagram for heart attack, high blood pressure, body mass index, and people

In order to develop the visualization stories, we applied an iterative development process of using data visualization for exploring and finding relevant correlations. Additionally, we enriched the interactivity with storytelling aspects resulting in a guided user experience.[10]

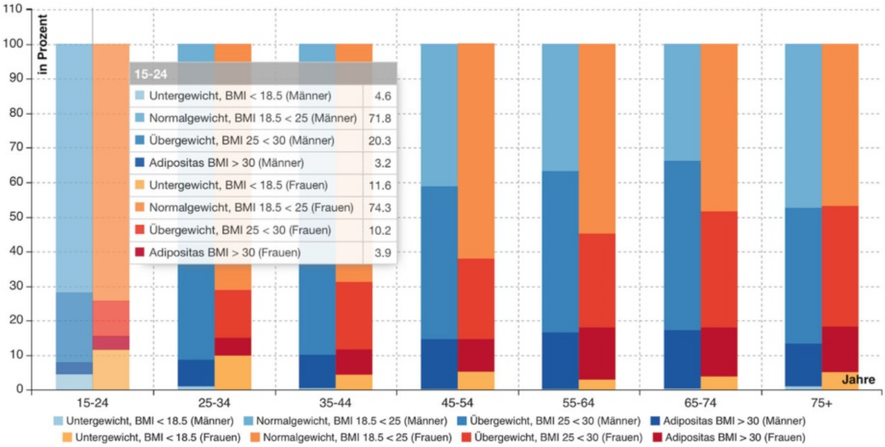
In order to meaningfully interpret visualized data, it is helpful to present chosen information in a way that comparison between some grouping and/or within a historical, time-dependent context is supported. Figure 14.8 presents the differences in body mass indices between age groups. The users can hover over the visualization and get more detailed information. For instance, users might be interested in the body mass index of people in the age range of 15–24. By clicking on the respective charts, the details about the color coding of the charts are presented.

This kind of visualization makes sure that the charts are not overloaded with information and gives the user the flexibility of zooming into details interactively.

Once the user clicks on the next page of the visualization, the story continues and informs the user about mortality rates. Figure 14.9 shows the historical development of mortality reasons of Swiss citizens over the last years based on different causes such as lung cancer or car accidents. The charts show that lung cancer was the main cause of death followed by suicide. For both causes, however, we see that the mortality rates were highest around 1980 and dropped afterward. Moreover, the users can use time slides to analyze concrete time periods in more detail.

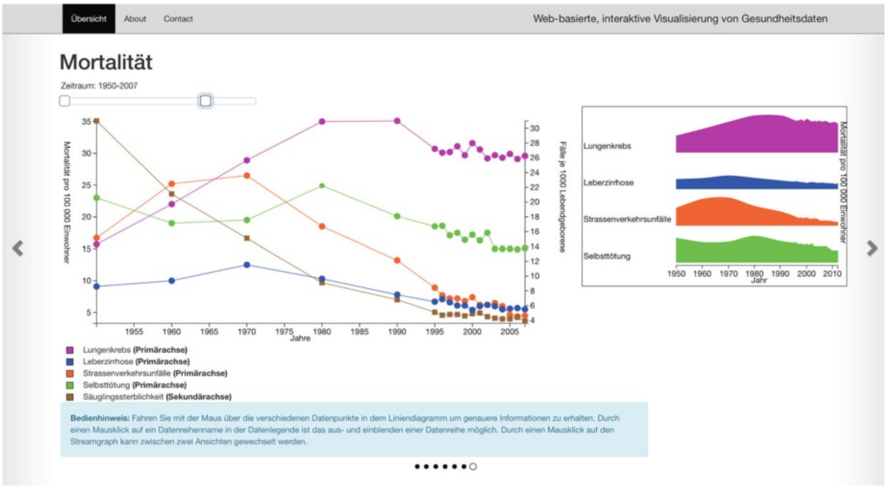
The next part of the visualization story provides more information on geographical differences about treatments in hospitals. The rationale of the narrative is that besides temporal comparison, regional differences are often of interest and may be

[10] http://www.visualcomputinglab.ch/healthvis/



15-24	
Untergewicht, BMI < 18.5 (Männer)	4.6
Normalgewicht, BMI 18.5 < 25 (Männer)	71.8
Übergewicht, BMI 25 < 30 (Männer)	20.3
Adipositas BMI > 30 (Männer)	3.2
Untergewicht, BMI < 18.5 (Frauen)	11.6
Normalgewicht, BMI 18.5 < 25 (Frauen)	74.3
Übergewicht, BMI 25 < 30 (Frauen)	10.2
Adipositas BMI > 30 (Frauen)	3.9

Fig. 14.8 Comparison of body mass indices between different age groups



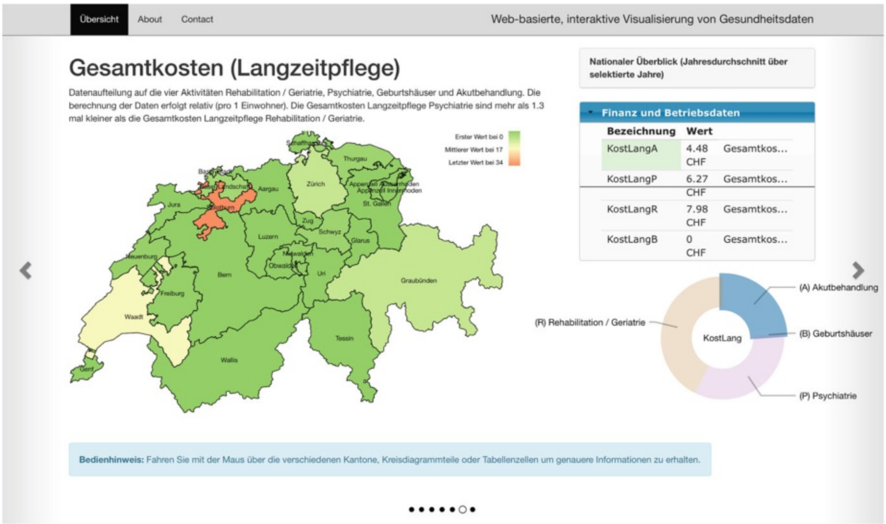
Übersicht
About
Contact
Web-basierte, interaktive Visualisierung von Gesundheitsdaten
Mortalität
Zeitraum: 1950-2007
Mortalität pro 100 000 Einwohner
35
30
25
20
15
10
5
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Jahre
Fälle je 1000 Lebendgeborene
30
28
26
24
22
20
18
16
14
12
10
8
6
4
Lungenkrebs (Primärachse)
Leberzirrhose (Primärachse)
Strassenverkehrsunfälle (Primärachse)
Selbsttötung (Primärachse)
Säuglingssterblichkeit (Sekundärachse)
Lungenkrebs
Leberzirrhose
Strassenverkehrsunfälle
Selbsttötung
1950
1960
1970
1980
1990
2000
2010
Jahr
Mortalität pro 100 000 Einwohner

Bedienhinweis: Fahren Sie mit der Maus über die verschiedenen Datenpunkte in dem Liniendiagramm um genauere Informationen zu erhalten. Durch einen Mausklick auf ein Datenreihenname in der Datenlegende ist das aus- und einblenden einer Datenreihe möglich. Durch einen Mausklick auf den Streamgraph kann zwischen zwei Ansichten gewechselt werden.

Fig. 14.9 Historical development of mortality reasons

presented in geographical maps. Figure 14.10 shows geographical differences of health care costs at hospital between Swiss cantons in a map of Switzerland. The user can interactively compare costs of geriatric, ambulant, psychiatric, and obstetric health care services. In addition, the visualizations are narrated with background information on hospital treatments as well as on interpretations of the results.

In Fig. 14.10 you see some textual description about the chart (the narrative) as well as geographical and tabular information. All visualization modes are interactive and interlinked and animate the users to explore the information from different visual perspectives.



Gesamtkosten (Langzeitpflege)

Datenaufteilung auf die vier Aktivitäten Rehabilitation / Geriatrie, Psychiatrie, Geburtshäuser und Akutbehandlung. Die berechnung der Daten erfolgt relativ (pro 1 Einwohner). Die Gesamtkosten Langzeitpflege Psychiatrie sind mehr als 1.3 mal kleiner als die Gesamtkosten Langzeitpflege Rehabilitation / Geriatrie.

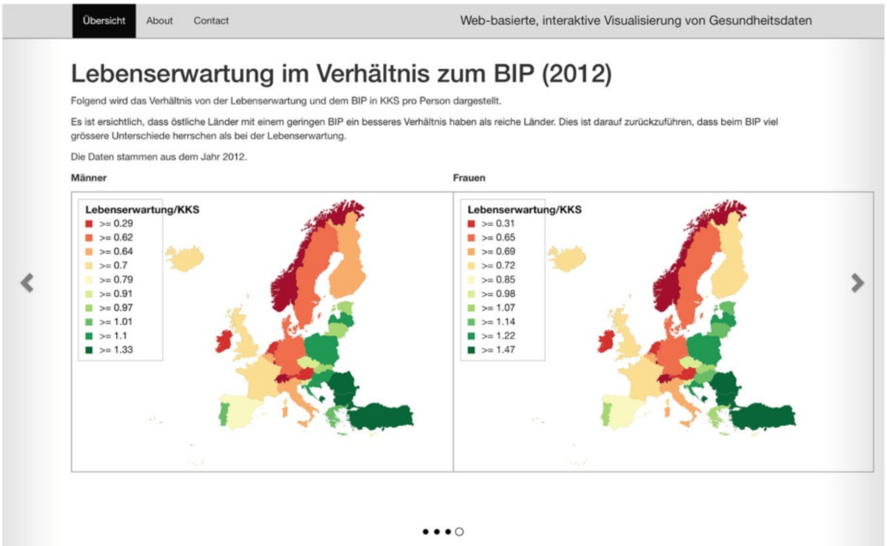
Nationaler Überblick (Jahresdurchschnitt über selektierte Jahre)

Finanz und Betriebsdaten

Bezeichnung	Wert	
KostLangA	4.48 CHF	Gesamtkos...
KostLangP	6.27 CHF	Gesamtkos...
KostLangR	7.98 CHF	Gesamtkos...
KostLangB	0 CHF	Gesamtkos...

Bedienhinweis: Fahren Sie mit der Maus über die verschiedenen Kantone, Kreisdiagrammteile oder Tabellenzellen um genauere Informationen zu erhalten.

Fig. 14.10 Regional differences of health care costs within Swiss cantons



Lebenserwartung im Verhältnis zum BIP (2012)

Folgend wird das Verhältnis von der Lebenserwartung und dem BIP in KKS pro Person dargestellt.

Es ist ersichtlich, dass östliche Länder mit einem geringen BIP ein besseres Verhältnis haben als reiche Länder. Dies ist darauf zurückzuführen, dass beim BIP viel grössere Unterschiede herrschen als bei der Lebenserwartung.

Die Daten stammen aus dem Jahr 2012.

Fig. 14.11 Correlation of life expectancy and gross national product between men and women within EU countries

The next step of the narrative visualization is to analyze life expectancy on a European scale. Figure 14.11 combines regional comparison and gender differences by presenting two maps of the EU side by side. The sequence of visualization guides the user through the examination whether there is clear evidence of a positive correlation between gross domestic product (GDP) and life expectancy. The results show that the ratio of life expectancy to the gross domestic product is better in eastern European countries than in western European ones—the main exception being Portugal. The reason is that differences in GDP are much higher than the differences in life expectancy.

The interested reader can experiment with more narrative visualizations at the following web page: http://www.visualcomputinglab.ch/healthvis/europeStory.html.

6 Lessons Learned

In the following section, we summarize the major lessons learned about developing interactive information visualizations based on open data:

- **Data preparation:** The most time-consuming aspect of information visualization is the data preparation phase, which follows similar principles as the ones used for building a data warehouse. Even though there is a vast amount of open data available, the datasets are typically loosely coupled collections of data items. Moreover, the datasets have very heterogeneous data formats and often lack a description of metadata. Hence, before data can be visualized, it often needs to be manually transformed, harmonized, cleaned, and brought into a common data model that allows easy visualization.
- **Visualization technology**: Recently, there has been a vast amount of visualization methods developed as part of the D3.js framework that enables quick prototyping. However, in order to develop consistent information visualizations, the produced charts often need to be adopted, for instance, to match dynamic scaling of axes, coherent color coding, and to enable persuasive interactivity. Hence, the high-level visualization frameworks that enable quick prototyping often cannot be used out of the box. In order to get full visualization flexibility, low-level visualization functionality needs to be customized, which requires writing much more code for achieving similar results. As a consequence, interactive information visualization and especially narrative visualization often require a development path from rapid prototyping using "out-of-the-box" data graphics toward "customized" visualizations that require some design and coding efforts.

7 Conclusions

The information visualizations shown in this chapter exemplify the benefit of gaining insight via interactive graphical presentations using open data available from public health authorities. In general, our society demands more transparency

(Taylor and Kelsey 2016). Open data coupled with information visualization will increasingly become an attractive way to communicate fact-based interrelations in economic, social, and political contexts. The awareness is increasing that it is important to provide open data to the public. Even so, it is relevant that the growing information available as open data becomes accessible via interactive visualizations in order to manage the growing complexity we are confronted with. Data scientists use their expertise in applying existing tools to process the vast amount of open data and to create interactive exploration environments for interested users and engaged citizens.

References

Bauer, F., & Kaltenböck, M. (2011). *Linked open data: The essentials*. Vienna: Edition Mono/ Monochrom.

Beeley, C. (2016). *Web application development with R using shiny* (2nd ed.). Packt Publishing.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '11), 17*(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185.7.

Huijboom, N., & Van den Broek, T. (2011). Open data: An international comparison of strategies. *European Journal of ePractice, 12*(1), 4–16.

Montana, M., & Zahner, D. (2015). *Web-based, interactive visualization of health data*. Bachelor Thesis, Zurich University of Applied Sciences, Winterthur, Switzerland.

Orszag, P. (2009). *Open government directive*. https://www.whitehouse.gov/open/documents/open-government-directive

Rosenfeld, L., & Morville, P. (2015). *Information architecture – For the web and beyond* (4th ed.). O'Reilly Media.

Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics, 16*(6), 1139–1148.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages* (pp. 336–343). Washington: IEEE Computer Society Press.

Taylor, R., & Kelsey, T. (2016) *Transparency and the open society: Practical lessons for effective policy*. Policy Press at University of Bristol.

Ware, C. (2012). *Information visualization – Perception for design* (3rd ed.). Morgan Kaufman.

Waser, F. (2016). *How can data analysis be fun? Visualization concept for databases*. Bachelor Thesis, Zurich University of Applied Sciences, Winterthur.