Serving official statistics visually

by Michael Neutze

Federal Statistical Office of Germany (Destatis)
65180 Wiesbaden
GERMANY

michael.neutze@destatis.de
+49 611 75-2410

Abstract

Destatis has been publishing sophisticated interactive data visualisations for far more than a decade. They all share the display of detailed data and users can understand this data much better while interacting with it. Among those visualisations are population pyramids, thematic maps, Voronoi diagrams and calendar visualisations. All have been custom built in-house and go beyond the confines of readily available chart types.

Users of regional statistics don't necessarily need to be GIS experts, nor need demography experts be able to program an animated population pyramid. But all those users would love to use such data visualisations in their work, no matter if it is of scientific, journalistic or educational nature. We therefore like to consider our sophisticated data visualisations as a service for data consumption, especially if we can promote embedding them into third party websites.

The presentation will give an overview of recently published visualisations, how they came into being, what kind of expertise was involved and how they were promoted. Special emphasis will be given to their embedding: What are the prerequisites that users want to embed our graphics, what are the requirements of our users and how do we retain our branding.

Keywords

Data Visualisation, d3, programming, embedding

1 Data Visualisation at Destatis

For almost two decades data visualisation has been in focus at the Federal Statistical Office of Germany (Destatis). While the necessity of it has never been questioned, a variety of efforts had to be established to get to the status quo. During the print era uniformity had always been the goal, e.g. standards for text sizes, line thickness and use of colour. Later on modern day knowledge of visual communication combined with the idea of a corporate design became prevalent. For a while choosing and learning to use the appropriate software packages were a significant part of the graphics department. Finally regular teaching classes for the effective use of statistical diagrams were added to the curriculum of further training.

In the meantime statistical dissemination could break away from the static paper medium to the screen and in fact Destatis published an animated population pyramid as early as 1999 as part of a CD-ROM based “multimedia project”.

Between 2003 and 2013 various interactive visualisations were published on the *destatis.de* website from animated population pyramids, to index calculators and many more. However these faced severe technical hurdles as users had to install plugins or were to use specific brands and versions of a browser, which users in larger organizations typically can’t do.

Fortunately in recent years the technology baseline in terms of internet connectivity and software capabilities is such that all visual user groups can be addressed.

There is now a dedicated section on our website where all the interactive visualisations can be found in one place (additionally to their place within the editorial hierarchy) and which is advertised directly on the homepage “above the fold”:

destatis.de/EN/Service/InteractiveVisual/InteractiveVisualised.html

Since not all visualisations are available in English (mostly maps aren’t) it is also advisable to have a look at the German language version of said overview:

destatis.de/DE/Service/InteraktivAnschaulich/InteraktivAnschaulich.html

1.1 Interactivity and Animation

In the beginning the Internet was just a transport or delivery medium for text and (statistical) data that later on was enriched by other media types. Nowadays video, directions and virtual or augmented reality are taken for granted thanks to location aware mobile devices. In between a lot of terms were used to indicate progress of the online medium, e.g. “multimedia” and “interactivity” come to mind.

Just like on paper, Destatis uses static diagrams (mostly in PNG format) to enrich its press releases and other online publications. Usually these help understand the data better and quicker but they are hardly ever being re-used by media outlets or other user groups and therefore should not be considered a service in the sense of this paper. Most users likely have enough expertise to re-create simple static diagrams themselves if only to adhere to their own design standards (fonts, colour).

Interactivity on the other hand has a broad range of meanings, from selecting different aspects of data, to revealing detailed information, from navigating a story broken into pieces via a stepper to sliding through an animation. The true meaning would be that a user could influence parameters of a statistic, e.g. a weighting pattern or the assumptions of a population projection. To put it more succinct with the words of Santiago Ortiz: “for a visualization to be interactive, you should be able to interact with the data not the image” [1].

However interacting with the data isn’t something statisticians are very keen for others to do to their data as they don’t want to convey the feeling of arbitrariness in choosing statistical parameters. So far only the index calculator for a personal inflation rate fulfils this aspect of interactivity. With it lay persons are able to calculate their ‘personal inflation rate’ according to the weighting pattern of their own consumption habits.

Typically the term ‘interactive’ is overused when it comes to data visualisation on the web and it is necessary to explain what kind of data visualisations are offered on the Destatis website and which of those could be considered as a service for data consumption.

Nevertheless those lesser forms of interactivity have proven pretty popular and useful. Offering basic thematic maps that allow the retrieval of data for a specific region is a standard use-case that delivers on the Shneiderman mantra of “Overview first, zoom and filter, then details-on-demand” [2].

Finally, animation can be used for a better understanding of statistics [3]. This is used to great effect in the population pyramid that has been published with every population projection for Germany since 2003.

1.2 Sophisticated data visualisations

Since terminology does matter, for the purpose of this paper the author suggests to talk about *sophisticated data visualisations*, comprising all sorts of the interactive spectrum and animation, but having in common that they are not static, that they are data-driven (i.e. changing the data immediately changes the visualisation) and they are not based on a template provided by a software product that's also in use elsewhere.

To make these data visualisations a reality, the only feasible way was to create them for the web right from the start. Therefore their evolvement was tied to the emergence of a standard for data driven vector graphics on the web, called SVG (Scalable Vector Graphics).

What this means is that the graphical presentation is a direct result of the underlying data, which offers a wide range of interaction possibilities, most prominently mouse-overs that reveal the data-point of a bar, line, pie or map region. Moreover the data visualisation can be easily updated just by adding fresh data to it.

1.3 Programming skills are necessary

None of the visualisations that are mentioned here have been produced with off the shelf software. Instead they have been programmed individually and in close collaboration with subject matter statisticians and their needs. It cannot be overstated how limiting the design process can be when one is confined to the available templates. With regards to PowerPoint some even argue that the template designs limit thinking and can be dangerous [4].

One might consider programming statistical diagrams from the ground up to be time consuming and costly and in the beginning this was really the case. But with the advent of D3 (*d3js.org*), a JavaScript library that provides all the necessary building blocks and tools for a data-driven diagram, productivity got a huge boost and the quality of the visualisations increased dramatically.

In recent years D3 has become the defacto language for data visualisation on the web – and as we will see below – for a lot of printed designs as well. It surely demands a different mix of skills, e.g. programming for the web *and* visual thinking combined of course with a deep understanding of statistics. But statisticians these days use programmes like SPSS, SAS or R to do their job and programming is always involved with these tools. Additionally no matter what level of interactivity or animation you are deploying in your data visualisation, it will always touch fundamental principles of software usability.

1.4 Deriving print graphics from the web

There are several technological developments that have been making the screen more similar to the printed page, namely the steep increase in screen resolution ("high-dpi" or "retina") and secondly that the web browser has become so much more powerful in the area of Scalable Vector Graphics (SVG). Until recently the printed page had had a much higher resolution than screens and statistical diagrams were prepared in vector format to achieve smooth curves and sharp lines in printed publications.

Since 2003 we have been advocating for the SVG standard mostly for its openness and scripting ability, meaning that the geometric forms could be manipulated programmatically in accordance to the change of statistical data and would furthermore offer all kinds of interactivity. In the decade after its introduction it became more and more clear, that those SVG files would not only be manipulated programmatically but that they would be created programmatically. The D3 language is a prime example for this. With D3 the basic workflow is to load data and then let D3 create as many line segments, bars, columns or map regions as the loaded data has statistical observations.

These days there is a vibrant community developing and sharing new data visualisations that are all done programmatically and would be very expensive or impossible to re-create using drawing and diagram tools like Adobe Illustrator or Microsoft Excel. Among these are population pyramids, calendar visualisations, chord- or Sankey diagrams and a variety of map projections never before possible. It usually doesn't take long after the publication of a successful and never before seen data visualisation that newspapers ask for a file of it that they can print. For a while this has been a challenge as everyone knows who has printed out webpages with disappointing results, mainly mushy or pixelated graphics and only the text being sharp.

It was just a logical step once technology had advanced enough to try to extract the vector graphic directly from the browser and finish it from there in design software suitable for print graphics such as Illustrator. Fittingly the New York Times, that for some years employed the creator of D3, has made available an easy to use tool for this extraction, aptly named "SVG Crowbar" (*nytimes.github.io/svg-crowbar*).

This workflow is not only convenient but it changes the meaning of an "online first strategy". Now not only can facts and figures be published first online because the medium is faster than print, even the visualisation can be created first for the web and a print version is derived from it afterwards.

1.5 Mobile Devices

The standard desktop monitor of the 21st century office worker has become much closer in size and quality to the printed page than one would have thought only years ago. While this might solve some of the restrictions of visualisation on the web – at least when thinking from the print perspective – at the same time new challenges in the form of tiny mobile screens have emerged.

More people access our information via mobile devices every day and the volume cannot be ignored. However adapting to smaller screen sizes works well for text which can reflow in different sized containers but diagrams not so much.

Among innovators or the digital native there is the claim "mobile or it didn't happen" meaning that information that cannot be consumed on mobile devices will likely be ignored. This however has to be taken with a grain of salt and the users of official statistics are still more likely to work on a desktop computer than most. But for journalists it may already be a close call.

What does this mean in terms of providing visualisation as a service? Mobile adaptations of data visualisations are a lot of work and usually require decisions of what to focus on and which aspects to leave out when catering for the small screen. But when done right the information can reach a much wider audience than ever before.

Especially with respect to the sharing of information on social media only few users will save a link to look at it later when they have a desktop size computer available. At least they need to get an idea if the information is relevant and then to share it.

So called responsive design, meaning the almost fluid adaptation of a website to different screen sizes and input methods, is a lot of work and becomes especially tough when data visualisations are concerned. But as with everything rare and difficult, it offers those who can master it a pole position in the market. And official statistics know their data best and first thus having a huge advantage against other information providers.

2 The competitive landscape of data visualisation

Earlier we mentioned a helper tool provided by the New York Times and their role in nurturing the D3 language. Indeed there are several news organisations which have advanced the sophistication of data visualisation and they have outpaced statistical organizations by a lot.

But news organizations still struggle with their business models and mostly those that serve a worldwide market can invest in innovative data visualisations.

Nevertheless there are exceptions, e.g. Berlin-based regional newspaper "Morgenpost" (*morgenpost.de/interaktiv*) has been winning many prices for their interactive stories. On the other hand especially in markets outside the English language there are respected quality news organisations that don't develop their own interactive data visualisations. They might e.g. embed interactive content like election results from a third party news service. Another approach is using template driven tools like *datawrapper.de* (available in several languages) that are popular among journalists to directly publish simple diagrams on the web.

The main observation is that there are users of statistical information that are capable and eager to use sophisticated data visualisations but don't have the means to produce these visualisations themselves. Constraining reasons might be time, money and expertise. In fact a reputation analysis commissioned by Destatis found that journalists of online media (including data-journalists) wish Destatis would provide more sophisticated interactive graphics.

Overall this is not a bad position to be in: Demand is high, expectations are high and suppliers are manageable. And as can be seen with regional newspaper "Morgenpost" investment is mainly tied to attracting the right kind of talent.

3 Visualisation as a service for data consumption

Data visualisations can become a service for data consumption if and when the following criteria are met:

- The quantitative information in question merits to be visualised, e.g. it is a large enough number of figures that would otherwise be hard to grasp.
- The appropriate visualisation for the data is not easily produced with off the shelf software.
- Producing the visualisation needs preparation in advance (e.g. maps)
- Long-time experience with the collection process of the data and its analysis are required to extract interesting or surprising findings in the data.
- The visualisation can be embedded in other websites or processed with the users' software.

Basically there is a need for data visualisation that the users won't produce for themselves due to a variety of reasons. Thus the NSI has a chance to provide data visualisation as a service, mainly in the form of embedding into websites.

3.1 Embedding

Embedding of a data visualisation means that our interactive diagram will be hosted on the *destatis.de* website but will be displayed as part of a third party website. To better understand the implications let’s look at video embedding for an analogy. A huge part of the success of video on the web lies in the embedding capabilities of services like YouTube or Vimeo to name just a few. Serving video on a webpage isn’t an easy task. It may cost a lot in bandwidth fees, different devices demand different video formats and ideally the quality of the video should adapt to the available bandwidth. To provide this all by yourself would be a lot of overhead to your regular website business.

But the web has this huge advantage that not all elements of a page need to be served from the provider of the website itself. The example of embedded YouTube videos in various websites is a testament to this. Even a school or a scientific researcher can now embed video in their blog with no additional effort and a lot of people do exactly that. In short, if an NSI could become the YouTube of statistical diagrams, that would be the ultimate service.

Fortunately the reality isn’t that far-off. Already Destatis has seen its animated population pyramid and interactive maps being embedded in websites of national newspapers. We are certainly still in the early stages of this process but have already learnt a lot. Not all news organisations will allow embedded content in their webpages, as they obviously lose control over that external content. They give away usage statistics of that page and also must rely on the provider of the external content to be powerful enough to sustain the load. Many news sites will have several orders of magnitude more page views than those of statistical offices. Typically one would discuss this plan with the people who run the webserver infrastructure of an organisation before advertising the embed capabilities.

There are plenty of sites that are glad to embed external diagrams from us. However their biggest constraint is their layout, which is in no small part a result of screen real estate reserved for advertising. Quite to the contrary the Destatis animated population pyramid has always made use of a full screen with its latest incarnation making good use of an ample 1280x800 resolution. On the other hand news sites may only devote their main column at about 600px in width for content. Destatis has used various techniques to cater for these users’ demands:

The population pyramid offers two kinds of layout. Besides the aforementioned layout in landscape orientation there is a mobile phone layout in portrait mode but with a slightly reduced interaction model. It is this mobile view that we provide for embedding.

Another time we created an interactive map that we did embed into our own website for technical reasons. The Destatis website allows for a more spacious layout than news sites but the map can determine itself programmatically if it is embedded in the *destatis.de* website or that of a third party and change its layout accordingly.

One of the risks in providing data visualisations as a service in the form of embedding is that readers of a web page may not be able to distinguish where different elements of said page originate. Retaining one’s own branding is therefore key. Again we can learn a little bit from the YouTube example. YouTube has a very strong branding with the red coloured play button in the centre of each video. That’s why it won’t be overlooked and the expectation to find the embedded video on YouTube standalone is never called into question. Similarly our embeddable map doesn’t only adjust its layout when embedded in a third party website, it also displays prominently the Destatis branding once being used outside of the destatis.de website.

As the usage of statistical diagrams in third party websites is still in its infancy the fact that we are happy to provide data visualisations as a service needs more publicity. This can either be done by using the standard web practice of the **</>** symbol for embedding but it furthermore requires talking to news organisations in advance of a publications to make them aware of this service.

3.2 Web Map Service (WMS)

Last but not least it is worth mentioning a visualisation service that is aimed at expert users of Geographical Information Systems (GIS). The German census of 2011 for the first time ever published selected indicators for the 1km-grid according to the INSPIRE standard. This data can only be interpreted in a useful manner if being visualised together with other map layers, e.g. roads, waterways or the building structure. To facilitate using these indicator an atlas of census results on the grid level was published at *atlas.zensus2011.de* (German only). Besides using those maps in the browser with some pre-selected thematic layers the census indicators are also available as a Web Map Service (WMS). This means GIS users can use those grid-level maps directly within their mapping projects and save a lot of time in doing so. This is another example that ease of use should make a good case for official statistics.

4 References

[1] Santiago Ortiz (moebio), “for a visualization to be interactive, you should be able to interact with the data not the image”. 14. August 2016 [Twitter Post], retrieved from *twitter.com/moebio/status/764588238950531072*

[2] Ben Shneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the IEEE Symposium on Visual Languages, pages 336-343, Washington. IEEE Computer Society Press, 1996. *citeseer.ist.psu.edu/409647.html*

[3] J. Heer and G. G. Robertson, Animated Transitions in Statistical Data Graphics. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 13(6), 1240–1247, 2007. *vis.stanford.edu/papers/animated-transitions*

[4] Edward Tufte, The Cognitive Style of Power Point: Pitching Out Corrupts Within, Graphics Press, 2004.