

Statistical Data

The census of statistical data has a long tradition. For centuries, population data have been collected and surveyed. Thus, administrations and agencies being authorised to collect statistical data are playing an important role. There are relevant institutions in almost every country in the world, sometimes just concerned with certain regions, carrying out this important public and social task. But also non-governmental organisations, associations and research facilities survey statistical data. Last but not least every corporation collects a multitude of data in their CRM and ERP systems being relevant for control, organisation and decision making. Furthermore, statistical data have enormous significance for decisions on issues in politics and economy. In medicine and social studies they are the basis for research and scientific work, for example in epidemiology.

For most people, access to statistical data is being offered by the media. Today, newspapers, magazines, or online publications can hardly do without displays of statistical data in the form of tables or diagrams – an increasing trend over the last 20 years.

Information Visualisation – Graphical Display of Statistical Data

Visualising larger amounts of data in tables is not very effective. Human perception works a lot better in recognising trends and distributions, if data are displayed graphically. The scientific discipline being concerned with the graphical preparation of numerical data is called information visualisation. The most common graphic displays are thematic maps, bar charts, pie charts and line charts representing time series. Lesser known graphic renditions are correlation or distribution graphics. Even though these graphical representations are very expressive, they are rarely used by the media, whereas so-called information graphics, being mainly combinations of the classical pie, line or bar charts in connection with images and editorial remarks, can be found very often.

Existing Software

There are several special applications available which can be used for analysis and further processing of statistical data that are mainly employed in a scientific environment. In corporations, this task is carried out by CRM or ERP systems working with integrated reporting modules. For single or isolated data, spreadsheet programs offer a fair possibility for quick display and analysis. However, all these systems have in common that they are usually no Internet applications. Even if visualisation features are offered, it is only possible to publish any results on the Internet via the export of static images. There are different approaches for publishing statistical data on the Internet: smaller suppliers often put documents on their websites in which tables and graphics are combined with editorial articles. The statistical offices in Germany offer a web front-end called GENESIS Online which allows for displaying and downloading individual tables in the form of a HTML application providing some graphic features. Eurostat offers a huge download portal for tables and documents which is also HTML based. In the U.S. the situation is even more heterogeneous. Every office maintains its own Internet portal where the information often needs to be collected from miscellaneous locations which are spread all over the portal structure.

Some technology suppliers have focused on software solutions which can be used for distributing data on the Internet. Some statistical offices worldwide apply these technologies for the purpose of distributed use and analysis of centrally collected data. In this connection, applications like Beyond 20/20 and PX-Axis can be mentioned. However, through their HTML user interface, these programs are limited in usability and functional range.

Furthermore, there are technology suppliers having specialised in certain graphic notations. The French software supplier Geoclip, for example, offers a Flash based map framework allowing to visualise data. This approach guarantees for good interactivity, but is rather based on maps and not on data.

State-of-the-art interactive Information Visualisation with DataDiver

Today, modern computers allow for a completely new approach what data visualisation is concerned. While, for example, in printed media one is limited to displaying single images, the drawing of a diagram or map can easily be done on a computer by selecting the corresponding features of the application program, which can, of course, also be a web-based application program. If the drawing process is then delegated to the client system, this can be done without the latency of a server query, allowing to manipulate or modify visualisations in realtime. In practice, this means that users can visualise and animate the variation of data over a certain period of time simply by sliding a control on the screen. Navigating through data in a contextual manner is also possible, for example, if the user wants to switch from one graphic notation to another one, or a similar indicator shall be used for another region. In the end, this functionality was eponymous for DataDiver, since all these contemporary options could consequently be put into practice with this application. Therefore, it was also critical to design DataDiver as a browser based application, being the only way to let users navigate through enormous data stocks in realtime and location-independent, change parameters spontaneously, and create interactive graphical visualisations for the evaluated data as quickly and easy as possible. And this is one of the main reasons why DataDiver does not only revolutionise the perception of statistical data, but also the way of making statistical data available, analysable, visible and usable for other purposes.

Many Data Sources – One Portal

Besides functional aspects, a centralised access was one of the central ideas when designing DataDiver. A portal should be built, being able to spontaneously deliver the demanded data to almost any kind of question. In order to make this happen, multiple and high-performance search options had to be available as well as a smooth transition from search result to visualisation by a maximum of three mouse

clicks. These key demands were met in an ideal manner by the keyword search which is easy to use in addition to the possibility to search data by means of a thematic catalogue or tree view. A portal offering answers to as many questions as possible requires a huge amount of data from all different kinds of sources. By taking over the data from the statistical offices in Germany and from Eurostat, a high coverage is already achieved for Germany and Europe right from the start. But also worldwide data are available through the data contingents of the World Factbook or the PENN World Tables and other international authorities. By successively expanding the database, also detailed data from other regions or areas in the world become available right in the starting version of DataDiver and will continuously be complemented, e.g. in the fields of demographics, health, justice, criminal statistics, elections, labour market and many others. This is guaranteed by constantly updating the database giving DataDiver full up-to-dateness with the portals of the original suppliers.

Further Application of Visualisations

A portal simply allowing the visualisation of data is not enough. What is the use of data mining and data visualisation if the results and images cannot be exported, embedded on a website, or used in online or offline presentations? This is the reason why DataDiver offers a multitude of export features. Besides print features, runtime versions of individual visualisations can be downloaded and embedded on a homepage or Intranet structure. If there is no Internet available in a meeting, for example, an offline runtime version can be created running on every notebook or desktop computer without being connected to the DataDiver server. Our hotlink functionality allows the embedding of DataDiver visualisations into a website simply by cutting and pasting a HTML code snippet. The visualisation will then directly be loaded from the DataDiver server.

Internationalisation – Consequently Implemented

DataDiver was built in Germany but designed for worldwide use. Not only different language versions are supported, but also different regional settings or locales. During login, regional settings can be selected DataDiver Background Information which configure the formatting of dates and numbers. Language and regional setting can also be set when exporting visualisations or presentations. For the launch, DataDiver will be available as German and English versions. For the French and Spanish versions work is already in progress.

Technical Implementation – Administration of Metadata

After having formulated the goals for the platform, it became clear very quickly that only a separation of the metadata from the live system would meet the requirements. An administration system had to be created which meets the following requirements:

- administration of metadata, such as indicators, variables and characteristics
- localisation tools
- description and classification of metadata by means of visualisation and calculation features being available on the live system
- description of formats for the original data to be imported
- generation, allocation and indexing of search items
- compilation of data for the live system from the original data and the metadata description, making it possible to have different compilations according to data sources and languages in order to generate
- runtime data sets for varying platforms
- description of conversion rules for compiling the original data
- multiuser application

The administration software was implemented as a generic Windows client communicating with a PostgreSQL database.

Technical Implementation - Client

When deciding on the client technology, besides the basic requirement of creating an application running in every browser, two other aspects were important and had to be observed. On the one hand, the features necessary for interactive data visualisation had to be implemented. Above all, this required the availability of a high performance graphics engine. On the other hand, a GUI framework should be implemented which could be used for the search and administration features in DataDiver. However, this framework should also be used for the quick implementation of other projects, such as CMS systems. At the same time, the appearance of the applications should correspond to those of a regular modern desktop application. Another criterion was easy maintainability.

A customary HTML interface had to be rejected simply because of lacking graphic features on the client side. To delegate the drawing and rendering processes to the server would have meant to accept weaknesses in interactivity and performance. The alternative of having an HTML interface with embedded SVG code by using ECMAScript on the client side did not lead to the desired results in performance and maintainability during testing.

This left the option to use a technology running on a virtual machine on the client side, reducing the choice down to Java and Flash. Java had the advantage of being an established technology offering a multitude of good development tools and integrated generic libraries for graphic and GUI functionality. Apart from that, Java support for Windows platforms is not very good and installing the Java VM is not really easy for inexperienced users. Flash on the other hand, is widely spread and offers excellent graphic performance, especially for the purpose of visualising data as vector graphics. Adobe's Flex technology and the open source framework OpenLaszlo would have been available as development tools. The decision was made in favour of Flash, since its broad acceptance and the easy installation of the Flash player, often even being preinstalled in many browsers, outbalanced the advantages of Java. Also, neither the frameworks Flex nor OpenLaszlo could

convince us with respect to the requirements we had.

This is why we developed two different frameworks. One framework to abstract the required drawing features needed to implement the interactive visualisations, and the second framework for implementing a window management which classic GUI elements, like buttons, lists, tabs, etc. This framework was then optimised in order to allow for any other application to be put into practice having the quality of a desktop application in very short time, such as CRM systems for example – however, as an application running in a browser. Another challenge was server communication when visualising data, making it necessary to extract any packs of data from them in extremely short time. In addition, these data volumes require persistence on the client side. The classical way of transitioning XML files into client objects was satisfactory with respect to the performance of random access, thus, very memory-intensive. This is why we implemented our own memory management for Flash being optimised in terms of parameters, transfer and processing speed, as well as memory consumption. We chose the open source compiler MTASC for compiling the Flash Action Script source codes.

Perspective

Already for the launch of DataDiver planned for Q4/2010, we offer a huge database. The basis of approximately 1 billion data fields sums up to a number of visualisations which cannot be quantified when taking all the different calculation and selection operations into account. Besides permanently updating the data stock, we will also constantly add new data sources and extend existing ones. In doing so, a strong focus will also be set on smaller suppliers whose data are of great interest for the public. We will also enhance the range of data for certain regions. But also the range of functionality will be enhanced with the aim to make DataDiver a universal visualisation framework for the Internet. In the near future, everybody will be able to visualise his or her own data with DataDiver, for example data which accrue in an organisation, business or in research. Using simple import options, like

CSV files or spreadsheets which allow for own runtime versions to be generated for websites, the Internet or offline presentations. The data can then be combined with data from DataDiver, standardisations, portions or quota can be calculated, and, as matter of course, the complete range of internationalisation features will also be available.