Dr Hans-Jörg Schulz is a computer graphics engineer researching novel methods for the integrated visualisation of hierarchically structured and heterogeneous datasets. Here, he talks about his inspirations, the challenges he has overcome and his hopes for the future.

How did you become involved in research on multiple heterogeneous data sources? What is it about this area that inspires you?

Before starting this project, I spent much time developing network visualisations and tree layouts. As it turns out, many graph visualisation methods can likewise be applied to the scenario of interlinked datasets. After all, these can be seen as a big graph themselves. What baffles and inspires me in these times of data abundance is the mismatch between the obvious need to put data in context by combining multiple data sources and the reluctance to do so. Like most software, visualisation systems still tend to work with one dataset at a time – a paradigm that is embodied in the File-Open dialog box for choosing a single data source. It’s time to change that!

What challenges do highly interlinked conglomerates of multiple datasets, or a ‘data landscape’, pose?

The main challenge posed by such a scenario is that the various data sources are heterogeneous. This means that they can have very different kinds of content: numerical data, documents, pictures, etc. Of course, these sources also require very different ways of visualising them; for example, plots for numerical data, tag clouds for documents and image galleries for pictures. Combining these with each other and providing a consistent interaction for an integrated visual analysis of such diverse data is a hard problem that’s far from being solved.

You have visually encoded data on various levels of granularity. What marks this advance as unique?

If data sources provide data of different granularities, it can be worthwhile to factor this into the analysis. For example, the analyst may want to investigate how data sources with state-level data differ from country-level data sources. Visualising data on different levels of granularity is a complex undertaking. That’s why most approaches show only one level of granularity at a time. With our visualisation approach, all granularities can be shown at the same time. While this idea is not new, I believe we were the first to apply it to multiple data sources on different levels of granularity. It’s one example of how we adapt methods for displaying graphs to the scenario of multiple linked data sources.

Could you explain what the concept of in situ visualisation is? What advantages does it have compared to other techniques?

The in situ visualisation principle provides a way to locally embed other related data sources, possibly also stemming from different levels of granularity, into a base visualisation of a current data source. We achieve this by using and extending the portal metaphor that turns a rectangular selection into a ‘window’ into a different dataset. Its main advantage is its intuitiveness, as it shows the required information from a different dataset right in the place where the analyst expects and needs it, instead of displaying it in a separate window on the side, as is the usual practice.

Has collaboration played a key part in your work thus far? Which disciplines are represented by members of your team and partners?

Visualisation is an applied science that would be nothing without any data from the real world to be visualised. Over the years, I have partnered mainly with experts from the life sciences, such as biologists and medical professionals. But I’m also maintaining very interesting collaborations with researchers from the fields of climate impact research, social networking, manufacturing, etc.

I love digging into different fields, learning something new along the way, and helping to solve the analytical challenges that arise in these very different domains with novel visualisation methods.

Looking forward, what will be your next steps in this project?

We are currently investigating a number of different research questions to answer in future work. To me, the most exciting of them is how to combine various individual visualisation applications instead of combining individual visualisation methods. This would take the whole scenario to the next level, aiming to develop solutions to break the current one-application-at-a-time paradigm by bringing together multiple applications under a common customisable hood. We already have a number of ideas on how to do this, but it will be a whole new research project in itself.
ADVANCES IN DIGITAL technology, computing power and global interconnectivity have all meant that there are increasing volumes of data being generated and collected concerning our lives and the state of the world around us. However, without the ability to manipulate, analyse, visualise and interpret these large quantities of information, the data are all but meaningless. Depending on the size and complexity of the dataset, there are many existing software packages that have been engineered to aid the task of data analysis with the aim of extracting useful information, suggesting conclusions and guiding decision-making processes.

The practice of data visualisation is closely related to data analysis in that it uses a schematic form to visually represent the data in some way that helps effectively communicate the information contained within. For successful data visualisation, it is important to consider both form and function of the graphical display so that it is interesting and engaging to interact with but also effectively conveys the required information.

Data visualisation is a growing field of investigation because of its usefulness in an environment where increasing amounts of information that we seek is contained within – to the layperson – impenetrable datasets. There are now many software packages designed to perform data analysis and visualisation on single datasets across a wide range of sectors and industries. However, data analysis is rarely that simple and to efficiently extract the most comprehensive information, it is becoming increasingly necessary to crosscheck between datasets and make use of additional information from alternative data sources.

Schulz and his team continuously work on multiple, simultaneous projects with the underlying theme of visual support for the analysis of hierarchically structured data from multiple heterogeneous data sources.

Using current software packages for data visualisation, this is a very laborious process, requiring much back-and-forth and switching between independent visualisations. To overcome this inefficiency, researchers are developing multimodal methods, whereby visual representations of multiple inputs are integrated from independent data sources in display. This is the context behind research being led by Dr Hans-Jörg Schulz at the University of Rostock, Germany. Schulz and his team from the Computer Graphics Research Group are working to develop methods that can help to visualise datasets in an integrated manner from a variety of different data sources. Furthermore, they consider multimodal visualisations that are heterogeneous in nature and have differential hierarchical structures and contrasting granularities.

THE NEED FOR A FLEXIBLE SYSTEM

To efficiently analyse data, it is important that the system being used is versatile because an analyst will not know, when starting to consider the data, where the information will lead. It may be that there are other data sources required for consideration and comparison to complement the data being analysed, or there may be a more appropriate data source that will be more suitable as the task evolves. To enable this, it is important that the relevant data sources are in a format such that they are flexible for inclusion in a visualisation and can be easily accessed.

It can become difficult to achieve this versatility when the different sources of data have heterogeneous forms, eg. pictures, written documents, numerical data, geographic coordinates, etc. This is an area of development that the researchers are working to improve, as Schulz describes: “The issue of incorporating
multiple datasets in a visual analysis and the challenges this implies will surely gain increased attention as more and more data are made openly available”. Consequently, there will be increasing demand to use visualisation techniques such as those developed by Schulz which will also encourage more scientists from other groups to build on this base to develop novel approaches for data visualisation.

Inherent flexibility is not only essential to incorporate and manipulate a data source; depending on the analyst, and the type of display that the data are being analysed on, the choice of strategy for data visualisation needs to be flexible and adaptable too, as Schulz explains: “For most of the resulting variations, suitable visualisation techniques already exist, but these are often unknown to analysts and not incorporated into common visual analysis systems”.

To address this issue, Schulz maintains a website called treevis.net which is a bibliography of existing and established tree visualisation techniques useful for representing hierarchically structured data. This platform enables data analysts to develop ideas for their own visualisation techniques and make use of those that have already been developed, as well as providing a representation of the scope of what tree-based visualisation strategies can achieve. Computer scientists in data visualisation are working toward developing similar reference libraries for the visualisation of other types of data sources; for example, survey.timeviz.net hosts a collection of techniques useful for visualising data that varies over time.

A SUCCESS STORY IN MULTIMODAL DATA VISUALISATION

Software solutions for data visualisation are meaningless without their application, using real-world data for practical data analysis purposes. A recent example of this is the development of StratomeX - a data visualisation system that can work with multiple stratified datasets. The package has been designed to be easily accessible to a broad base of disciplines by requiring no use of scripting language and provides a straightforward and scalable overview of different component groups across different data types. The project is a joint research effort together with Graz University of Technology and Johannes Kepler University of Linz in Austria; and Harvard Medical School, USA.

A specific application that the StratomeX software is being used for is in the identification and characterisation of cancer subtypes. An understanding of this is necessary for refining therapeutic targeting and consequently improving patient outcomes. This software, enabling the integrated analysis of different molecular data types, can be used to search for and identify specific subtypes of cancer for which tailored therapeutic approaches may be appropriate. “It works by clustering various data sources from patients suffering from the same type of cancer,” Schulz explains. “We then visualise these different clusters and, by considering how they correspond in each of the data sources, biomedical experts are able to derive specific cancer subtypes.”

Patient samples are grouped according to, for example, DNA or mRNA methylation profiles, copy number levels of a particular gene or gene mutation status. These groups are then represented as columns and the clusters within them are represented as components of these columns. Visualising complex data in this way enables relationships to be more easily identified between neighbouring columns. The resulting ability to diagnose a particular cancer subtype is highly beneficial, as they often have different underlying molecular causes and thus will respond most effectively to differential treatment plans.

MULTITASKING

Schulz and his team continuously work on multiple, simultaneous projects with the underlying theme of visual support for the analysis of hierarchically structured data from multiple heterogeneous data sources. Having made significant progress in this area to the benefit of many data analysts and scientists needing to visualise the interactions between datasets, Schulz is keen to take the possibilities of data visualisation to the next level. This will involve the development of a map-like interface that can be used to visualise large, heterogeneous data landscapes to not only map out the data, but also incorporate a number of levels of data granularity linked to visualisation and analysis techniques, and provide efficient mechanisms for analysts to navigate through this visual multilevel data map.

HANS-JÖRG SCHULZ received his Master’s (2004) and PhD degree (2010) from the University of Rostock, where he is currently working on a research project on the visualisation of heterogeneous data. Schulz frequently applies his research in domains such as biomedicine and systems biology. In his free time, he is an avid hiker and climber.