Peter Kunszt is familiar with the challenges that accompany research involving big data.

Peter Kunszt manages SyBIT, the SystemsX.ch bioinformatics and IT project

“SyBIT gets scientists in shape for big data”

In systems biology as elsewhere, big data has long been recognized as an invaluable source of information. Yet important insights into biological systems are only conceivable if it is possible to extract the relevant information from the enormous volume of available data. Peter Kunszt and his team support researchers in automating such processes. Through the IT support project SyBIT, they help Swiss scientists keep pace with the world leaders in the long term.

Where does the data flood in systems biology come from?

Research in the field of systems biology only became possible after significant advances had been made in the technologies required to monitor biological systems. Considerable developments have been achieved, for example in DNA sequencing equipment. The progress in mass spectrometry and, most recently, microscopy techniques is also impressive. All these devices produce increasingly large amounts of data. This situation can be compared to progressively more powerful digital cameras; new models offering more megapixels are released each year. These call for new storage devices with more storage capacity. The same is true for research, but with much bigger dimensions, which is why we call it big data.

What does the concept of big data encompass for you?

Particularly in systems biology, big data is not only characterized by huge volumes of information but also by its complexity. Often, we have to deal with different types of values and must first determine how they are linked. Such data is not easy to interpret. Another aspect is speed. Large volumes of data are increasingly difficult to analyze within a reasonable period of time. Data uncertainty, i.e. the quality and reliability of the information, is an additional factor, as values can be inaccurate due to measuring errors.

Which one of these aspects poses the greatest challenge?

Volume, complexity and speed always go hand in hand. It is our job to rapidly find the best solutions to scientific problems together with the investigators, using the technology available today while taking these three factors into account.

How exactly does SyBIT support SystemsX.ch scientists?

Every project is different. Depending on the needs of the investigators, we help them assemble the required hardware and software or we arrange for access to mainframe computers. But we also help the scientists analyze, interpret, manage and store their data.

Through the SyBIT project, we place the whole know-how related to data management at the disposal of the researchers, and help them automate their processes and make them efficient. In a manner of speaking, we get the scientists in shape for big data so they can fully exploit the potential of new technologies.

Christa Smith © Frank Brüderli

Has the need for IT support increased?

Yes, definitely. As the volume and complexity of the data increases, it becomes more and more difficult and time-consuming for the investigators to manage their data. You might think this is easy to do, but the recording of the data alone is complex when dealing with volumes such as those generated by mass spectrometry, for example. The data stemming from different experiments must be correctly annotated and filed in a structured manner. This is essential so that the information can later be assigned to the correct projects and, if necessary, be reproduced. The analysis and the assessment of the huge data volumes call for algorithms that automatically pinpoint characteristics and patterns of interest.

Can you mention a specific example?

We are currently supporting MorphogenetiX. In this project, the scientists are studying cell specialization using 3-D microscopy. Thanks to this new technology, the samples no longer need to be formalin-fixed, but can be filmed live. The 3-D microscope takes up to 700 pictures per second, allowing the researchers to witness cell division and ultimately demonstrate how a cell such as a specialized brain cell comes into being.

The data volume generated by this procedure is enormous. SyBIT is helping the MorphogenetiX project team with the computer-assisted processing of the data. One of my collaborators is on site for several months; he is testing the developed algorithms alongside the project's specialists in order to automate the processing of the vast amount of data.

The sheer amount of data calls for high storage capacity. What is stored and what is rejected?

We need to understand the data extremely well to be able to decide what is relevant. This is why, in today's data-heavy research, it is of utmost importance to first grasp the meaning of the data and to detect patterns and correlations. At the beginning of a project, especially in basic research, the meaning behind the data is often not understood, which is why scientists usually want to hold on to it all. Often it only becomes clear towards the end of a project which data are relevant and which data can be deleted, due to the fact that they can be easily and even more precisely reproduced at a later time.

And how do you make sure the data is still accessible in the future?

Unfortunately, long-term data storage is still a largely unsolved problem. In the fields of genomics and proteomics, international central databases have already been established, but long-term financing is not yet guaranteed. No solution is yet available for the archiving of data generated by imaging technology. Once SyBIT comes to an end, no institution will take over the management of this data.

Who, in your opinion, needs to take on responsibility for this matter?

In my view, data archiving is the libraries' job. Scientists should not have to pay for the storage of their data. The government must find solutions. Luckily, the problem has been identified, and several options are currently being tested and discussed at the political level.

SyBIT and SystemsX.ch will both come to an end in 2018. Who will guarantee IT support thereafter?

Local support groups. The idea for these originated in the Lake Geneva region. The Vital-IT group was founded there as early as in 2004. This organization offers computing power, memory and support in the area of bioinformatics. We were able to establish local SyBIT partners at the Universities of Zurich and Basel, as well as at the ETH Zurich based on this model. By doing so, support for the scientists is guaranteed beyond the duration of SyBIT and SystemsX.ch.

And how can this know-how be anchored within the community?

Fortunately, we are already seeing SystemsX.ch investigators applying their newly acquired skills to new projects. The SyBIT specialists do not exclusively support SystemsX.ch projects, so the acquired competences are introduced into other research groups as well.

Are Swiss universities now prepared for data-intensive research after SyBIT?

Yes, in principle. The local support groups offering bioinformatics services are firmly established in the universities and the SIB Swiss Institute of Bioinformatics will handle their coordination after SyBIT expires. Over the past few years, we have also helped the SystemsX.ch partner institutions develop the required IT infrastructure. The task now is to establish connections between these local IT resources, so that the universities can take advantage of all the available services and specialized infrastructure. This will help Swiss scientists remain among the global leaders in systems biology research in the long term.