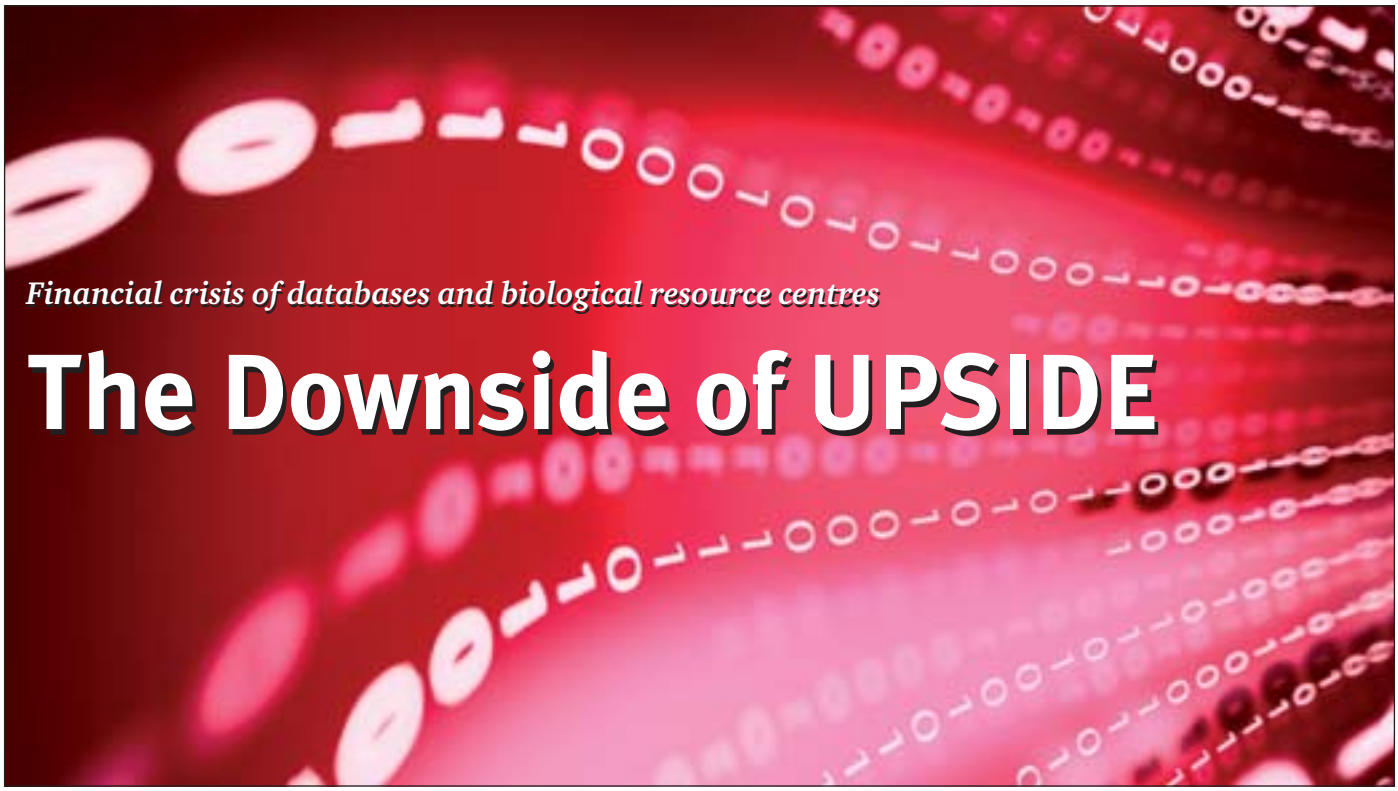

Financial crisis of databases and biological resource centres

The Downside of UPSIDE

Infrastructure is not only essential for urban planning. With the ever-growing flow of huge amounts of data that need to be managed, it becomes more and more important in research, too. The only question is: who's going to pay the bill?

An African proverb says, "Knowledge is like a garden, if it isn't cultivated it cannot be harvested." Scientific research naturally generates a lot of knowledge, just think of modern biology approaches, like whole genome sequencing, cumulating in a data deluge of biblical proportions. However, in order for science to advance and enjoy the fruits of success, all this data in line with the Uniform Principle for Sharing Integral Data and materials Expeditiously (UPSIDE) needs to be stored, managed – cultivated – and made available free of charge to the entire scientific world from Vladivostok to Tierra del Fuego. Databases do exactly this job.

No problems in the beginning

In recent years, they have become more and more the staff of life for every self-respecting scientist but only a few are aware that a lot of effort and thus money is needed to run and maintain this modern day knowledge collector. Usually, money is not an issue to worry about when a database or a biological resource centre is set up that archives and distributes scientifically relevant organisms. In many cases they are financed together with the project they arise from or they receive some initial funding when the collected data is of high relevance to the scientific community. However, this initial funding only lasts for a limited amount of time, three to five years in general, after which most databases and resource centres have to fear for their existence. The reason for this being that, at present, national funding agencies only support the developmental phase as they don't yet have any policy to finance established, long-term research infrastructures.

There are, of course, good reasons and – what's more – an increasing need to do so. With every new "big biology" project – like, for example, the recently announced 10K genome project aiming to sequence the genomes of not less than 10,000 vertebrate species – the amount of data to be managed grows inexorably. It's a kind of bottomless pit as new technologies will even facilitate and speed up the generation of ever more data in the future.

To give you some idea of what dimension we're talking about here: the nucleotide sequence of the whole human genome would take up three gigabytes of computer storage but those are only peanuts compared to the data storage at the European Bioinformatic Institute, EBI, which hosts databases such as the EMBL Bank (DNA sequences), UniProt (protein sequences), Ensembl (animal genomes), Protein Databank (3D structures), ArrayExpress (gene expression) and IntACt (protein-protein interaction). In September 2009, the EBI recorded a histori-

Well-sorted is halfway found.

cal storage of 4.5 petabytes of data, which converts to as much as 4,500,000 gigabytes and it won't take long before we enter the realm of exabytes, incredible 10^{18} bytes.

Funding won't last forever

However, the exponential growth in data is not yet accompanied by an exponential growth in funding. On the contrary, monetary contributions from governmental funding agencies that also extend beyond the developmental phase are urgently needed in order to be able to update, maintain and integrate all the data on a regular basis. At the same time, it must be made easily accessible to the data consumers, researchers worldwide. It's actually a rather small price to pay when one considers the vast amounts of money that were spent to create the data in the first place. The costs for the "Human Genome Project", for example, totalled an estimated €3 billion but only a fraction of that is needed to curate all the data obtained.

Back in 2005, *Nature* already looked into this issue with the special report "Databases in Peril". Out of 89 databases operating in 2000, only 18 said they had no financial problems five years later, 51 were struggling financially and seven had to shut down because their funding had ceased. (*Nature* vol. 435: 1010-11).

And the situation doesn't seem to have improved much lately with TAIR, The Arabidopsis Information Resource, being the latest victim of national agencies' funding policies. Over the last ten years, TAIR has established itself as the first source for everyone involved in *Arabidopsis* research with the help of two consecutive five-year grants from the US-based National Science Foundation, NSF. But last year, the NSF decided to phase out the support of TAIR as they don't have any dedicated grants for long-term funding of research infrastructures. Currently, TAIR still has $1.6 million at their disposal, as much as they were awarded the five years before but, from September on, the budget will decrease every year until 2013 when funding by the NSF will inevitably cease. This is a disaster for TAIR as the NSF was the only supporter of the database. Suggestions were made to introduce subscription fees or find alternative funding sources, including grants awarded by countries other than the USA, but as promising as it sounds, there's always a catch that makes the suggested strategy not feasible (see interview with Eva Huala, director of TAIR, on page 18).

Possibe Business Models

Like TAIR, many other databases and resource centres face an uncertain future unless solutions to get them out of their financial dilemma are quickly found. And what better way to find a rapid solution than to get people, who suffered or still suffer from funding difficulties, into the same room to talk about it. Exactly this was done at the workshop "Database & Bioresource Sustainability Models" held in Rome, Italy last November. The meeting was organised by Casimir, a European Community (EC) funded initiative aimed at the co-ordination and integration of databases set up in support of projects within the EC Research Framework Programmes 5 and 6, and brought together scientists from different fields of research wrestling with the same problem – finding a stable, long-term financial resource.

One issue of the workshop was how to manage the transition from a not-for-profit company to a self-sustaining enterprise. Johannes Maurer, former scientific director of the German Science Centre for Genome Research (RZPD) and now managing director of imaGenes GmbH, embarked on exactly this course when his company's source of finance suddenly ran dry. Founded in 1995 within the framework of the German Human Genome Project, the RZPD received governmental funding until 2007, after which it was planned to continue as a for-profit business. "The idea obviously was that after some years of 'not-for-profit' work RZPD would in a miraculous way become self-sustaining. But self-sustainability in such a dynamic environment as the life sciences is far from being trivial. Even with RZPD's well established service branches the centre was heavily dependent on funding. Self-sustainability could only be reached by promoting healthy business units, while at the same time cutting loss-making parts. At RZPD we had all prerequisites to create such a healthy business when the funding stopped. However, becoming self-sustaining was never a complete changeover for us: for our business model, the integration in and close cooperation with the scientific community is key. In other words, being commercial and providing an open access is not in itself an antagonism – as long as 'open access' is not used synonymously 'free of charge'," explained Johannes Maurer about the difficulties in getting the business started.

However, if "open access" does mean "free of charge" then some business models are clearly not suitable. And thus, according to a recent article published by the newly established open-access journal Database, *The Journal of Biological Databases and Curation*, biological resource centres and data resources are more or less left with

Interview Eva Huala, TAIR

"Subscriptions Are Not a Good Idea"

When TAIR was launched in 2000 did you expect it to become that big and crucial for Arabidopsis *research?*

From the beginning we've had a very ambitious vision of what the database could become. Despite our success, I think there are still many things we could do to increase the power and usefulness of TAIR. In each funding cycle we've had to scale back our potentially most transformative ideas due to budget cuts.

What was the initial long term plan for financing the database?

We expected to follow the lead of other model organism databases such as SGD (for yeast), MGI (for mouse), FlyBase (for Drosophila), etc. that have received stable grant funding for decades from NIH. There seems to be a realisation by agencies funding animal research that cutting-edge databases are foundational to cutting-edge science, however, this realisation seems to be lacking on the plant side.

What immediate consequences will the decrease in funding have starting in the autumn of this year?

We've already made staff cuts with the result that less of the latest research on the function of *Arabidopsis* genes is being captured in TAIR. Because we haven't captured them, these recent experimental results are not available to researchers unless they scan the literature themselves, which means that they are unlikely to be used in genome-scale omics experiments or annotation of the latest plant genomes. Within the next few months, we'll be cutting back further to the point where new information is added only if it's submitted to us by researchers directly. Unless a significant source of new grant funding is found, TAIR will no longer exist in two to three years.

What alternative funding options are currently considered?

Although 75% of our usage comes from outside the United States, it's very difficult to arrange for TAIR funding from other countries due to restrictions on transfer of funding outside the country. Ironically, it's possible for investigators in Europe to get research funds from the United States for certain types of projects but not vice versa. We're continuing to apply for smaller grants from various US funding agencies and we've just launched a new corporate sponsorship programme that we hope will bring in a bit of money to keep us going in the short term. If we don't get much of a response from the sponsorship programme we may have to go to a mandatory subscription system for companies. I strongly believe that subscriptions for academic researchers are not a good idea. To make that work we would be forced to shut off the flow of data to other resources like Entrez Gene and Ensembl Plants, which would hurt the whole field. Also, we depend on the goodwill and sense of ownership from the community that motivates them to submit data. This goodwill could easily be damaged if we start charging for data that was freely sent to us.

In Europe, the importance of securing biological informatics data has been recognised and joint efforts funded by the EC were initiated. How's the situation in the US?

As I said above, it seems to depend on whether you work on plants or on an organism considered relevant to human health, such as flies, roundworms or yeast. For some reason, it hasn't yet been registered at the highest levels that the human population depends, even more critically, on both basic and applied plant science to achieve a sustainable future. We need plant science to advance at least as quickly as biomedical science and we need databases and other resources like those used for biomedical science, for that to be possible.

Should TAIR really have to shut down, what would that mean for Arabidopsis research?

According to the many researchers who have commented on our funding situation at our website (*http://arabidopsis.org/doc/about/tair_funding/410*), not just *Arabidopsis* research but plant research overall will be significantly held back without TAIR.

INTERVIEW: KATHLEEN GRANSALKE

only two not-for-profit options, "Cost Recovery" and "Institutional Funding" (Chandras et al., *Database* vol. 2009:bap017; doi:10.1093/database/bap017). According to the authors, "Institutional Funding" is the most promising model, however, as already mentioned there are still too few public institutions, which have dedicated funds set aside to support long term research infrastructures. In "Cost Recovery", the expenses of a project can be recovered fully or partially; however, only the partial cost recovery model, with the core funding being provided by one or several funding agencies and the marginal costs being recovered from end users, was found to be viable.

One biological resource centre that adopted this business model is EMMA, the European Mutant Mouse Archive. EMMA was set up more than ten years ago with financial help from the EU ringing in at €3.9 million. "But in 1999 the facility was hit by an EU framework funding ban and then sustained by its host institution at the CNR Institute of Cell Biology in Monterotondo", EMMA project manager Michael Hagn remembers. However, after "successfully securing grant funding in three different EC framework programmes in highly competitive calls", the non-profit repository, currently archiving over 1,000 mutant mouse lines, seems to have it all wrapped up. "EMMA operates on a partial cost recovery model and provides a free-of-charge archiving service, whereas the distribution of mouse resources is provided for a service fee that is used solely to meet the animal husbandry and stock replacement costs," Hagn explains. But the current business model might not provide enough financial resources for future needs, as according to Hagn, "The demand for the EMMA services has seen a continuous increase since the establishment of the repository and the gap between existing and required capacities is getting wider each year."

Preventing the worst

To stop the gap becoming an abyss, European politics need to step in and this has already happened. In 2002, the European Strategy Forum on Research Infrastructures (ESFRI) was launched in order to "identify and address the scientific needs of research infrastructures in Europe for the next 10-20 years." Projects from all sorts of research fields like Social Sciences and Humanities, Environmental Sciences, Energy, Materials and Analytical Facilities, Physical Sciences and Engineering as well as e-Infrastructures were selected and many of them have already entered the preparation phase. In the Biomedical and Life Sciences field the following projects made the cut:

- EATRIS – The European Advanced Translational Research Infrastructure in Medicine (*http://www.eatris.eu*),
- BBMRI – European Biobanking and Biomolecular Resources (*www.biobanks.eu),*
- INFRAFRONTIER – The European Infrastructure for phenotyping and archiving of model mammalian genomes (*http://www.infrafrontier.eu),*
- ECRIN-PPI – Infrastructures for Clinical Trials and Biotherapy (*www.ecrin.org),*

Home to the frozen mutants at EMMA

- INSTRUCT – Integrated Structural Biology Infrastructure (*http://www.instruct-fp7.eu),* and
- ELIXIR – Upgrade of European Bioinformatics Infrastructure (*http://www.elixir-europe.org*).

Among these selected projects, the ELIXIR initiative, which to-date involves 32 partners from 13 countries, looks like the magic potion databases and resource centres urgently need to cure their financial ailments as, according to their programme, one of ELIXIR's goals is to "secure funding commitments from government agencies, charities, industry and intergovernmental organisations". The project itself, launched in 2007, is funded by the European Commission Framework 7 Capacities Programme for Research Infrastructure and, if everything goes as planned, will enter its construction phase in 2011. Fortunately, ELIXIR doesn't have to start from scratch – it will be built upon the existing IT infrastructure at the EMBL-EBI located at Hinxton, UK, which will then become the "central hub of the European life science infrastructure". However, the super-modern technology comes at a hefty price; according to the 2008 Roadmap report by ESFRI, the construction costs for hardware and networking amount to a mind-boggling sum of about €500 million, let alone the estimated operational costs of €100 million per year.

A worthwile investment

But it seems that this is a price that many European countries are willing to pay because the first financial commitments to ELIXIR have already been booked to the project's account. The United Kingdom, or to be exact, the Biotechnology and Biological Sciences Research Council (BBSRC) was the first to contribute £10 million last year, followed by Denmark (€5 million from the Danish Agency for Science, Technology and Innovation and several Danish universities). Just recently, Finland also chipped in €1.85 million. The money is provided by the Ministry of Education through the Academy of Finland along with some co-financing by the Institute for Molecular Medicine Finland (FIMM), the IT Center for Science (CSC) and the National Institute for Health and Welfare (THL). It is intended for pilot studies in 2010 but additional financial support has already been promised. Olli Kallioniemi, director of FIMM, says about Finland's contribution to the pan-European project, "A small country should concentrate its investments in the European dimension on those areas that are of critical importance and where synergies may arise."

No more waste of resources

And that's a point well taken as in future, the data production rate will further increase with every new and cheaper technology that becomes available. Thus, the need for "well-oiled" research infrastructures becomes compelling and, in order to guarantee open access to those resources for everyone, governmental help is desperately needed. It can only be hoped that it won't come too late for some resources because, in the worst case scenario, when a database goes down, all data it contains will be lost. And apart from the fact that it can be a very expensive loss, knowledge is a terrible thing to waste. KATHLEEN GRANSALKE