European Commission

Horizon 2020
European Union funding
for Research & Innovation

Case studies on Open Science in the context of ERC projects – Set 2

February 2018

This document presents the second of five sets of case studies that have been produced in the framework of the *'Study on open access to publications and research data management and sharing within ERC projects'.*

European Research Council
Established by the European Commission

Contract number:	ERCEA/A1/2016/06
Contracting authority:	European Research Council Executive Agency (ERCEA)
Contractor:	PPMI Group UAB with University of Edinburgh and Georg-August-Universität Göttingen
Authors:	Angus Whyte, Digital Curation Centre, University of Edinburgh Birgit Schmidt, Georg-August-Universität Göttingen Lisa Matthias, Georg-August-Universität Göttingen Viltė Banelytė, Public Policy and Management Institute
Contact:	Dagmar Meyer, ERCEA
Contact email:	erc-open-access@ec.europa.eu
DOI:	https://doi.org/10.5281/zenodo.1235345

LEGAL NOTICE

These case studies have been prepared for the European Research Council Executive Agency (ERCEA). However, the views set out in this document are those of the authors and/or the interviewees only and do not necessarily reflect the official opinion of the ERCEA. The ERCEA does not guarantee the accuracy of the data included in this document. Neither the European Commission nor the ERCEA nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

TABLE OF CONTENTS

SUMMARIES

Harnessing Proto-Enzymes for Novel Catalytic Functions (CompEnzymeEvolution)

Lynn Kamerlin, a Professor at Uppsala University and the lead of the project CompEnzymeEvolution applies computational methodologies to study protein evolution and to design new catalytic functions. In this context, Professor Kamerlin highlights some of the benefits and challenges that her team has encountered when applying open science practices in their research environment, particularly related to managing research data and the costs of open access publishing.

Dynamics of Local Transcriptomes and Proteomes in Neurons (Neuronal Dynamics)

When Professor Erin Schuman received an ERC Advanced grant for her project Neuronal Dynamics, there were no formal requirements for PIs to publish articles in open access or share their research data openly. Still, most of the publications derived from the project have been published in open access and research data has been deposited in publicly accessible databases. Professor Schuman sees clear benefits that publishing in open access and opening research data could bring to science.

Neural circuits underlying complex brain function across animals - from conserved core concepts to specializations defining a species' identity (BrainInBrain)

The small field of insect neurophysiology has until now lacked a repository of openly shared data to help researchers collaborate on cross-species analysis of specific brain regions. But there is significant potential for such a repository to yield fresh understanding of decision-making, intentionality and navigation. Associate Professor Stanley Heinze, of Lund University, leads the BrainInBrain project, an ERC Starting Grant exploring how a region known as the central complex helps insects know their direction. Heinze explains the key role the *Insect Brain Database* plays in sharing and visualising images produced from the research. He describes how this resource helps researchers find and share the relevant knowledge to aid discovery, and his strategies to address the challenge of sustaining the resource for the longer term.

1. Harnessing Proto-Enzymes for Novel Catalytic Functions (CompEnzymeEvolution)

Summary

Lynn Kamerlin, a Professor at Uppsala University and the lead of the project CompEnzymeEvolution applies computational methodologies to study protein evolution and to design new catalytic functions. In this context, Professor Kamerlin highlights some of the benefits and challenges that her team has encountered when applying open science practices in their research environment, particularly related to managing research data and the costs of open access publishing.

1.1. Introduction

Lynn Kamerlin, Professor of Structural Biology at Uppsala University, received an ERC Starting Grant for her project *Harnessing Proto-Enzymes for Novel Catalytic Functions* (CompEnzymeEvolution) in 2012. Professor Kamerlin explains that enzymes are nature's catalysts, speeding up chemical reactions that drive life, from millions of years to seconds. As enzymes regulate almost all of life's processes, when enzymes go wrong, this leads to disease and death in extreme cases. Therefore, enzymes have been generating a lot of scientific interest and have been extensively studied from a biochemical and biomedical perspective. There is great scope also for using enzymes as biocatalysts for therapeutic and synthetic applications, bioremediation, and the generation of novel biofuels, to name but a few examples. The past few years have seen great progress in designing artificial enzymes, but the tremendous catalytic proficiencies of the best naturally occurring enzymes are, as yet, unmatched by any manmade system.

Kamerlin and her team are combination biologists who use advanced computational tools to understand the fundamental drivers of enzymes' evolution to gain new catalytic functions, and to develop and implement novel computational methodologies for artificial enzyme design. Her ERC project has contributed to resolving long-standing controversies around the mechanisms of one of the most fundamental reactions in biology. It illustrated how enzymes could functionally evolve at the most fundamental, molecular, level and provided a new toolkit, CADEE, for the computer-aided directed evolution of enzymes, which her team is currently using to manipulate enzyme function *in silico*.

1.2. Successful open science practices used in the project

- *What are the main outcomes of the project so far?*

The main outcomes of the project can be categorized into three different directions. The first is a fundamental computational methodology development. As mentioned above, this includes a new toolkit for performing computational directed evolution. Professor Kamerlin sees the benefit of open science in this regard as for software development the sharing of methodologies is very important.

The second outcome addresses several scientific controversies concerning the fundamental chemistry of the studied reactions, and fundamental questions on how proteins evolve in general. Kamerlin explains that in order to understand how enzymes function, it is important to first understand the fundamental chemistry they catalyse very well. Following this, in order to understand how to engineer enzymes for new catalytic functions, it is valuable to understand how they have evolved and what tricks nature uses to modify them.

Finally, the overarching goal of the project is to harness proto-enzymes (ancient ancestors of today's modern enzymes) for new catalytic functions. As one clear example of this, Kamerlin's team collaborated with colleagues at the University of Granada and the Georgia Institute of Technology who are leading experts in the reconstruction of ancestral proteins and their biophysical characterization. In this work, the team resurrected Precambrian β-lactamases, the precursors of modern antibiotic degrading enzymes (and thus key participants in the race against antibiotic resistance), and showed how with a simple substitution, these enzymes can be taught to perform completely novel chemistries. However, this modification *only* worked on ancient and not modern enzymes, highlighting the power of ancestral proteins as scaffolds for new chemistries.

To date, the project has produced 30 publications with several more in various stages of preparation and peer review, 2 software advances (a new toolkit CADEE, and a V6 upgrade on the Q software package[1]), as well as new methodology models in biology. All publications are open access (using gold open access options to ensure immediate access), both the CADEE and Q6 source code have been deposited to GitHub (a Git version control repository hosting service which is mostly used for computer code), and all research publications have extensive supporting information presenting the associated data sets. Finally, the research team recently moved to (and have had very positive experiences with) online data repositories, and they expect to make much more use of these for the project's remaining publications.

1.3. Challenges faced and success achieved

- *What challenges did you face in leading such an "open" project?*

Professor Kamerlin considers that the biggest challenge for her team, as for any computational group, are the costs: "There are great differences in publication frequency across disciplines, and computational (bio-)chemistry is on the higher end". Kamerlin's project has published 30 publications so far, amounting to over €50 000 in article processing charges (APCs). APCs vary widely from journal to journal based on perceived prestige. Since publication costs are eligible for reimbursement, the ERC grant made it possible for the project team to make all publications open access. However, Kamerlin remarks, "I consider myself very lucky to have had the funding from the ERC to be able to cover these costs. If I had only had a small research budget, I couldn't have afforded those APCs. It's an unfortunate situation that in terms of research evaluation people still publish in hybrid, the more 'prestigious', journals, where you need to publish if you want to get future grants. It's a serious problem how some journals are exploiting OA requirements".

Another challenge that Kamerlin's team encountered was the time needed for research data management activities. Data curation is time-consuming and it adds a lot of extra work on top of the normal project workload. However, it is also critical for research reproducibility. Kamerlin has experienced a lot of pushback from her students because they were "annoyed at how much extra time they had to spend preparing all this data for depositing in repositories", and it has been a challenge to build their understanding why this is important.

- *How do you deal with pushback from students?*

Although Professor Kamerlin has to train new groups of students and postdocs from point zero every year, and there is still some pushback, she sees the situation changing in a positive direction. She explains, "the first students really gave me a lot of pushback and I still hear many complaints and a lot

[1] Q6 Repository: EVB, FEP and LIE simulator, https://github.com/qusers/Q6; CADEE: Computer-Aided Directed Evolution of Enzymes, https://github.com/kamerlinlab/cadee

of mourning, but at least the new ones are starting to understand that there are certain things expected of them. They also see the other work that has been published that already has the data deposited and this has been really helpful". In the beginning of a new project, there usually is not much to build on. But now the students can use the extensive input and parameter files from previous projects, among other things, as a baseline and guide to understanding the methodology, and to better know how to approach their projects. Being able to refer to these (already existing) large appendices instead of (re-)creating them *de novo* every time also saves a lot of time, which the students appreciate.

Kamerlin observes that the younger generation is becoming more open to open science practices, but that they need specific skills and training. "This training needs to be incorporated in the curriculum, and open practices need to become a natural part of how science is done. That will help a lot with the pushback", she says.

- *Do you see any challenges or advantages for pre-print servers, where you can make your papers openly accessible for free?*

Professor Kamerlin has had great experiences with depositing her work in bioRxiv[2] and her team's software code on GitHub. The project team is currently facing a long article publication process, which is not uncommon for the Life Sciences. In this discipline, it can take more than a year to publish an article. One of the papers, currently under peer review, is already openly available on bioRxiv. Depositing a copy of the manuscript as a pre-publication made it immediately available and citable (since bioRxiv provides DOIs). Some pre-print critics dislike the fact that the manuscripts have not been peer-reviewed yet, but Kamerlin says "once you reach a certain stage in your career, you can use your own scientific judgement as well and you have experience to be able to make a qualified assessment. It is fantastic to have access to science as soon as it comes out, including work from the top researchers in the field, and not having to wait 12 months or more for the work to complete the peer review process."

Professor Kamerlin's research is at the crossroads of Biology and Chemistry and she is publishing in journals of both disciplines. She noticed that there are differences in publication practices and cultures in the two disciplines: "Elite journals in Biology will accept and encourage bioRxiv, but elite Chemistry journals actually consider this as a prior publication and do not accept manuscripts which have been posted to pre-print servers. This is gradually changing, albeit at a seemingly slower pace than in the Life Sciences." She explains that these policies pose a potential problem for the team if their more interdisciplinary work does not get accepted by their preferred Life Sciences journals and they then submit it to Chemistry journals instead. Since they have already deposited the article in bioRxiv, it might not get accepted by the publisher. Nevertheless, Kamerlin plans to submit more pre-prints to repositories in the future and hopes for "a culture change in Chemistry because these archives are really valuable".

- *You already said that open access publishing costs are very high, but what about data storage and preservation costs?*

Due to the large amounts of data generated, the project team had to find a cost-efficient and reliable solution for data preservation. Acknowledging the ERC grant for the financial support, Professor Kamerlin believes that "the most important thing is to find a company that we can trust (that we know it will not vanish in five years), that our data would be citable and the cost for storing data is not unreasonable, because we're in a really data-intensive field". Again, comparing disciplines, requirements and costs for data curation and storing can differ greatly. For some disciplines, data

[2] https://www.biorxiv.org/

means a few Excel spreadsheets, but for others, like Kamerlin's, it means hundreds of terabytes for an individual research team and high data depositing costs.

The team stores all the data on their server at the University of Uppsala, but also deposits a basic data package on Dryad[3], a curated data repository that hosts data underlying scientific and medical publications and makes them discoverable, freely reusable and citable. This package contains input files, basic files, a package of certain sample structures, explanations of what the input files are and what kind of parameters you need to repeat the simulations, in order to ensure reproducibility of research conducted by Professor Kamerlin and her team. The project team just recently started using the Dryad data repository. Before, their papers had 200-page appendices, most of it being raw data, but for the two most recent papers, they deposited the data packages in Dryad. This repository was chosen because "it's a great cost versus ease-of-use balance". However, Kamerlin "wouldn't put hundreds of terabytes of data into Dryad because the cost is prohibitive". The project team has their own data servers to store all data associated with publications. Anyone who wants to access the data can either get a guest account on request or send (or visit with) a hard drive and receive a copy of the requested data.

1.4. Impact of open science practices

- *Do your colleagues support open science?*

Professor Kamerlin says that many of her colleagues at her institution are not actively supporting or practicing open science. "Unfortunately, I work in an environment dominated by a very old-school way of thinking about how science is done", she explains. The extra effort involved in curating and managing data is a challenge for many people. They feel they could be working on another paper instead of curating the data. Kamerlin feels that it is going to be a lot of work to get people to see the benefits of data sharing: "I'm hoping once they understand the benefits, people will be more encouraged. I do have like-minded collaborators among the younger generation. Amongst my peers, I do see that there is an increased welcome for this," says Kamerlin. In addition, there is one senior colleague who is very supportive of open code and who Kamerlin refers to as an "open science champion". Through his position as an eminent computational chemist, he has also been a great help in showing Kamerlin's students that open practices are very important by engaging in conversations with them.

- *Where do you see the benefits of "open" practices?*

Seeing that open practices are not yet part of researcher evaluation, Professor Kamerlin does not see specific career benefits as such. However, she mentions two other major benefits of open science practices that are very important to her: transparency and inclusion. Biomedical research is witnessing big discussions about reproducibility, and this is also a huge problem in computational research. To foster reproducibility, it is important to Kamerlin to make as much underlying data accessible as possible. The two main software initiatives of the project, the Q simulation package (version 6, Q6) and CADEE, are both available on GitHub and are completely open source. Having the code openly accessible is very important. Kamerlin, quoting a colleague, says "because when you can't see the code, you can basically program the code to say print this number. If you can't really access and validate the software, you'll never know what people actually did." Hence, in order to make her work reproducible and verifiable, Kamerlin embraces transparency.

[3] http://datadryad.org/

She states, “I do it because I believe in it. And actually during the publishing process we have received compliments in referee reports about how great it is that we’re actually showing so much data.” It is appreciated by people who read the papers. It has also led to new collaborations for Professor Kamerlin because people could access the data and then were able to ask her questions. She says that “being open gives exposure to your science in a different way. If your paper is closed, people might read, cite, and think about it, but if things are open they can engage with it, rather than just read it and then put it on the shelf”.

The other aspect regarding openness that is very important to Professor Kamerlin is who is included in this discourse. In her opinion, people who are saying that open access is expensive and not seeing the value of it, often do not consider that the world of science is really large, and not everybody is equally privileged. Those in rich Western institutions are only a very small percentage of the total number of people doing science. Small institutions with lower budgets in large parts of the world often cannot afford the big subscription fees. Hence, Kamerlin thinks that it is important that *all* people who wish to work with the data have access to them, since “it leads to a much healthier exchange of ideas and collaboration. You share knowledge in a different way than just hiding it behind paywalls or burying your primary data”.

2. Dynamics of Local Transcriptomes and Proteomes in Neurons (Neuronal Dynamics)

Summary

When Professor Erin Schuman received an ERC Advanced grant for her project Neuronal Dynamics, there were no formal requirements for PIs to publish articles in open access or share their research data openly. Still, most of the publications derived from the project have been published in open access and research data have been deposited in publicly accessible databases. Professor Schuman sees clear benefits that publishing in open access and opening research data could bring to science.

2.1. Introduction

Brains are dynamic organs and can change their properties to encode the behavioural experiences of animals allowing them to adapt to future situations. The ability of the brain to respond adaptively relies on modifications to the existing proteins. Since many of the changes due to environmental stimuli occur at synapses (a point of communication between the nerve cells), the question arises as to how the modified synapses gain access to the new messenger ribonucleic acids (mRNAs) and proteins to maintain their functions. The main hypothesis of the Neuronal Dynamics project (*Dynamics of Local Transcriptomes and Proteomes in Neurons*) was that the ability of synapses to strengthen or weaken over time depends on the local pool of mRNAs and newly synthesized proteins.

The project was led by Professor Erin Schuman who is the Director of the Department of Synaptic Plasticity at the Max Planck Institute for Brain Research in Germany. The ERC Advanced Grant enabled Professor Schuman and her team to test the hypothesis and map out all the mRNAs present in the dendrites and axons (extensions of the nerve cells) and mRNAs present in the cell bodies of neurons. In addition, the team developed a technique that allows one to visualize a newly synthesized protein of interest in intact tissue, which was not possible before. These results closed a critical gap in the understanding of how different patterns of activity are read out by rapid modifications of sets of proteins produced near the synapse. In addition, these data provided one of the first "systems" views of the synapse.

2.2. Successful open science practices used in the project

The overarching goal of the project was to understand mRNA and protein populations present in the brain and more specifically in the hippocampus area. The team has successfully achieved both of these goals. The project resulted in a number of publications as well as datasets deposited in specific databases openly available to anyone.

- *The project was funded under Framework Programme 7 at the end of 2011 when no formal obligations in terms of publishing in open access applied to the ERC grantees. Still, the publications that derived from Neuronal Dynamics were published in open access. Why did you choose to make your articles openly accessible?*

"Most of the journals that we in my lab like to publish in are open access or at least partially open access journals", says Professor Schuman. In addition, the papers associated with the ERC grant were published in the same pool of journals that her team usually publishes in while working on other projects funded by other grants. Hence, choosing open access journals was a usual practice for Schuman's team in the Neuronal Dynamics project.

- *How do you choose the journals for your publications and do you consider their open access policy?*

Professor Schuman explains that when the research results materialise and are written up, the team starts thinking about the journals that they could publish in. The first aspect that is considered is the journal's impact factor: "We start thinking of the journals that have a high impact factor. I don't really want to say the highest impact, but practically speaking, we aim for the highest impact journals that we can publish in". The team also considers the open access policy of the journal and most of the time they are already aware of it, as the pool of journals they typically publish in is rather small: "We already know what the open access policies are [of most of the journals]. If I were to consider a new journal, then I might scrutinise their open access policy. But the majority of the new journals that we think about, such as eLife[4] for example, are designed in the open access era and are well aligned with the open access movement".

- *In your research field a lot of data are produced. Do you share the data openly?*

"Yes, we generate a lot of data and we upload it to public databases so that everybody can access it", says Professor Schuman. She adds that the data are uploaded to a database already during the review of an article, although initially they are closed and not available to everyone. The reviewer is provided with a password and can access the data for the paper under review. As soon as the article is published, the data are either unlocked or uploaded to a public database.

The research team have chosen specific databases that they use for depositing data. For example, the proteomic data is uploaded on the Pride (PRoteomics IDEntifications) database[5] and the Transcriptome data is stored in the NCBI Sequence Read Archive[6] (SRA) repository. Pride is a standard database for proteomic data and is particularly appreciated by Schuman and her team as it has an easy-to-use interface: "As soon as the article is accepted, one can simply push a button and the data become openly available for everybody", she explains.

2.3. Challenges faced and success achieved

- *Are open access and open data sharing common practices in neuroscience and is this something the researchers in this field are striving for?*

Professor Schuman thinks that sharing data openly together with a publication should be a common practice, although currently it is not always the case: "I review papers where the data are not accessible and there are cases when we would like to analyse the data ourselves, apply a different set of criteria or a different significance threshold. The data, however, are not available. Of course you can ask for them, but then it is not quite the same as just having the data there to be able to try things out. The data are shared by some researchers, but it is not a standard practice yet".

Regarding open access to publications, Schuman thinks that largely all of the top neuroscientists are aware of open access and are making their best efforts to implement it. There might be some differences in open access take-up between various disciplines. However, she does not see the discipline as the main hindrance for wide open access implementation. For her, the policies of publishers are more of an obstacle in that regard. Professor Schuman notes that some journals, such as the learned societies' journals, were faster in adopting open access policies than the big private publishers. The latter publishers have many popular and high impact journals across various

[4] https://elifesciences.org/

[5] https://www.ebi.ac.uk/pride/archive/

[6] https://www.ncbi.nlm.nih.gov/sra

disciplines that many researchers aim to have publications in. But those publishers are not interested in changing their policies unless funding agencies, such as the ERC and others, put pressure on them to modify their policies.

- *Did you face any obstacles when publishing your articles in open access during the ERC grant? Did you experience any financial, legal or other types of issues*?

Professor Schuman and her team did not face any real obstacles related to open access publishing in the Neuronal Dynamics project. "In general, there are some issues, such as paying for a publication twice to the publisher [paying article processing charges as well as a subscription fee, also called "double dipping"] that do not make me very happy. That does not seem right. But other than that, we did not have any issues with open access publishing", says Schuman. She also points out that her team usually publishes in journals that are high ranking but also provide open access to publications after a six-month embargo period and do not practice "double dipping".

- *What about publishing articles after the end of the ERC funding period? Did you face any financial problems in this regard?*

To a large extent the grants and funders that Professor Schuman works with, including the ERC, consider the publication costs as eligible. After the grant ends, article processing charges can be covered with her lab's running costs. However, as many journals that her team publishes in provide a free open access option after an embargo period, paying for open access is rarely needed.

2.4. Impact of open science practices

- *What is your motivation to practice open access and openly share your research data?*

Open data sharing implies additional efforts in terms of dedicating more resources, whether money, time or people who would curate the data. Still, Professor Schuman believes that sharing data is important for scientific progress: "For us, we have the added desire to share the data, as a lot of the data that we generate are large datasets and it would be very selfish and not good for science if we just kept the data for ourselves. By opening our data, we actually hope that other people will use them and analyse them in the ways that we would not analyse them ourselves, or might choose a subset of proteins or transcripts to analyse that we would never analyse or go into in detail. When you generate as much data as we do in our lab, there is no way that you could sufficiently study it by yourself. That kind of data should be made available to everybody, so that the information is shared, analysed and science can progress faster".

Schuman also notes that open data sharing can pose a risk. Researchers having access to the same data might choose to work on the same issues, which increases competition. She adds: "But the competition is good. It is a sign that you are working on an interesting question which is interesting to many people in the field".

3. Neural circuits underlying complex brain function across animals - from conserved core concepts to specializations defining a species' identity (BrainInBrain)

Summary

The small field of insect neurophysiology has until now lacked a repository of openly shared data to help researchers collaborate on cross-species analysis of specific brain regions. But there is significant potential for such a repository to yield fresh understanding of decision-making, intentionality and navigation. Associate Professor Stanley Heinze, of Lund University, leads the BrainInBrain project, an ERC Starting Grant exploring how a region known as the central complex helps insects know their direction. Heinze explains the key role the *Insect Brain Database* plays in sharing and visualising images produced from the research. He describes how this resource helps researchers find and share the relevant knowledge to aid discovery, and his strategies to address the challenge of sustaining the resource for the longer term.

3.1. Introduction

How do insects know where they are going? What can the flight of the bumble bee, or the deliberations of a dung beetle, reveal about our own decision making? An emerging scientific community is investigating how a small neurological structure, the central complex, may reveal answers to one of the key questions in neuroscience: how brains deal with intentions. To support their investigation, the community is collaborating openly to enhance analysis of data and resources across insect species.

Associate Professor Stanley Heinze is eminent in the insect neurophysiology field, one of eight Principal Investigators in the Vision Group at the University of Lund, Sweden. With the help of the ERC in the shape of a Starting Grant that began in January 2017, Heinze is developing the *Insect Brain Database*[7]. This resource is already working to support the main objective of the project *Neural circuits underlying complex brain function across animals - from conserved core concepts to specializations defining a species' identity*: understanding the workings of the central complex, the 'brain in the brain' of the project title. The research develops the core insight that this region of the tiny insect brain can serve as a model for larger brains with hundreds of thousands more neurons. The research may reveal how animal brains, including our own, encode intentions and decision-making in pursuit of those intentions.

3.2. Successful open science practices used in the project

The BrainInBrain project is looking specifically at how the behaviour of training insects correlates with their brain activity, taking account of their varying lifestyles and ecological niches. The challenges of this research are that it requires the integration of many varieties of data across a wide spectrum of species. The data includes multiple digital image types, including 3D images of individual neurons, the very high resolution and bulky files derived from confocal microscopy, and schematic representations of brains. Consistent labelling is needed to search across species and brain structures, and then

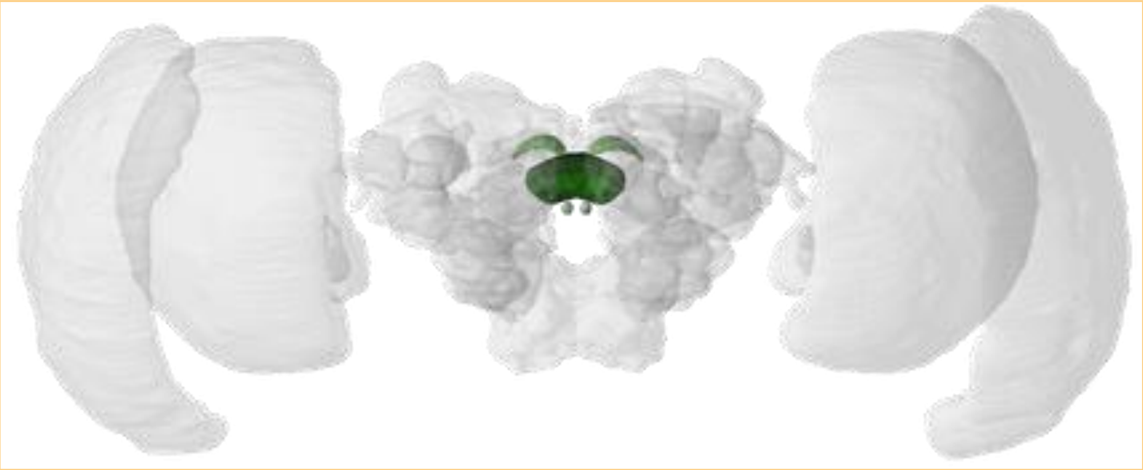
[7] https://www.insectbraindb.org/

analyse the results. Satisfying this need sets the *Insect Brain Database* apart from more mature data resources in other areas of neurobiology that model organisms for a single species, such as Drosophila. From the beginning, the *Insect Brain Database* has been developed with the goal of open sharing in mind.

Professor Heinze founded the *Insect Brain Database* in 2017, aiming to make it as useful and intuitively easy to use as possible for his research community. By late 2017, the initiative had gained the financial backing of seven neurophysiology labs, who each contribute to the shared pool of data. The ambition is to include most labs working in neurophysiology using non-classical arthropod model organisms.

This public database makes data from Professor Heinze's research and the contributing labs freely available, and archives published data. This potentially saves time in literature searching, as according to Heinze, "if you are interested in a brain region, say the lateral accessory lobes, finding out which cells have been recorded for that brain region across different species would take about a month of literature search… often in obscure journals and covering long periods." By contrast, the new database offers a simple graphical search engine and a graphical display of search results, within and across species. In doing so, it also reduces the anatomical knowledge needed to find the information.

Figure 1. Brain of a Monarch butterfly highlighting central complex. An example of an image from Stanley Heinze (2017), Insect Brain Database (https://www.insectbraindb.org/structures/5/5/)



© Stanley Heinze CC BY-NC 4.0

The *Insect Brain Database* also offers contributing laboratories a private working space to help researchers find and organise their data prior to publication. This is based on the realisation that good data management begins at home, and stems from Heinze's personal experience: "While doing my postdoc on the results of my PhD, one of the reviewers wanted to see one of the raw drawings that I had used and it took three days of digging around in wardrobes full of external hard drives to find it." The experience of struggling to find data on request was an incentive to develop something to help. It also gave Professor Heinze "… a wake-up call that we do need good data management plans. We need to know where data is stored and to use good naming schemes in a way that standardises the information you attach to files. Once you have that, it makes you very aware of the kind of data you have, and the data you actually need to obtain. Also, you need to consider whether the data you can obtain fits with the data you already have, and whether there is a justifiable need for new resources."

The database currently includes various data types produced in the neurophysiology research field. As well as photographs, there are 3D reconstructions of individual neurons and brains, plus schematic and semi-schematic representations of them. Transforming the raw data files for viewing online in web browsers took time to get right, but the software now renders the SWC files (source files for 3D reconstruction) into object files while uploading the data. "This made it easier for users to upload their data and integrate it automatically into the database. The *Insect Brain Database* also includes confocal image stacks which are large, up to a gigabyte in size, and therefore split into a compressed version for online viewing and a version that users can download to work with offline".

A new standard for labelling brains, the standard brain nomenclature,[8] has added a great deal to the utility of the database in Professor Heinze's view. Applying this standard across species and data types has enhanced the system's capability to organise these data to help researchers find, integrate, and reuse them in ways that will be compatible with other systems. "There was a long discussion across the insect brain and anatomy field for almost 10 years. That finally culminated in a common naming scheme for parts of the insect brain. This got unified just before we started, which was really useful".

- *How does the database operate?*

The data in the database is freely available for anyone to access without registration. In addition, anyone can register and on doing so gain the ability to submit new data (typically image files) and to download files. The quality assessment, organisation, and enhancement of the data are performed by a curator; and there is one curator for each of the species included in the system. "No one person can be expert for all species", explains Professor Heinze. Each curator is responsible for approving data for one species. Heinze himself is the head curator, training each species' curator in the system's functions.

The database is also developing a laboratory data management application. This aims to provide a workspace that labs can use to share data privately prior to publication. When ready for publication, data will get a persistent identifier such as Handle, or Digital Object Identifier (DOI). Data may also be submitted to a journal editorial platform for peer review and made public when the paper is accepted.

The ability to cross-link from a publication to the data will add value to the scholarly record. Typically, the 3D visualisations are submitted for publication as supplementary material in the form of video files. But cross-linking from the article to data in the *Insect Brain Database* will be more interactive. For example, readers will be able to link to a bibliography for each neuron cited in the article.

The license conditions that apply to research data can be an impediment to reuse. If the conditions are too restrictive or are not explicitly stated, then others may not be legally able to reuse it. To address this issue, Professor Heinze has adopted a path he believes will encourage reuse. The Creative Commons Non-Commercial license is the default option for new contributions. The contributors' lack of awareness of licensing options is an issue; "in my experience researchers don't usually know about this", he says.

- *When you won the ERC grant, you had an option to opt in voluntarily to the Open Research Data pilot. What was it that led you to participate in the Open Research Data Pilot?*

Beneficiaries of ERC grants funded under the ERC Work Programmes 2015 and 2016 may opt-in, on an individual and voluntary basis, to the Horizon 2020 Pilot on Open Research Data in order to facilitate access, re-use and preservation of research data generated during their research work. Participation carries certain obligations with it, which include preparing a Data Management Plan, although beneficiaries may opt out at any stage. Despite these obligations, Professor Heinze remains enthusiastic about opting in.

"Generally, I'm a big fan of sharing open resources" says Heinze, "and I think this comes from the fact that we're a very small field. Insect neurophysiology is not like medical neuroscience where there are

[8] Ito, K., Shinomiya, K., Ito, M. et al. (2014). A Systematic Nomenclature for the Insect Brain. *Neuron*, Volume 81, Issue 4, Pages 755-765. https://doi.org/10.1016/j.neuron.2013.12.017

hundreds of thousands of researchers; there are only a few of us." This community is dispersed across the globe. Until recently, the field has lacked the shared data resources that larger Life Sciences communities such as molecular biology have, with limited exceptions for the study of the fruit fly (Drosophila); e.g. Flybrain[9] and Virtual Fly Brain[10]. Consequently, it has lacked the efficiencies such resources can bring. Heinze grew aware of the need for a shared database during his time working on the Monarch butterfly. The ERC grant and the opportunity to work with a skilled software developer provided the motivation to get this initiative off the ground.

The data management plan (DMP) for BrainInBrain, which he submitted in keeping with the ERC guidelines, was based on his existing plans. Professor Heinze says the DMP is being updated to reflect plans for further data types, e.g. behavioural videos and electron microscopy files: "These are beyond our current capacity and haven't been implemented yet, but the DMP is being kept alive".

3.3. Challenges faced and success achieved

The *Insect Brain Database* is already proving a useful resource for its contributors to address the scientific challenges, which are the justification for the open database. It provides a central repository for digital image data on brain anatomy and physiology of different insect species. Facilitating open sharing promotes wider collaboration and helps make the case for sustaining the database. But there are further scientific drivers and challenges.

During his PhD research, Heinze took an interest in the central complex. This small group of cells operates like a compass, telling the insect the direction it is heading in relation to the sun. Heinze's work as a postdoctoral fellow found similar sun compass neurons in the Monarch butterfly. It transpired that migratory species always have this kind of mechanism in their brains. Hence, the key challenge that the database helps address is to understand whether non-migratory species of insects and other animals share this neuro-circuitry. The project is identifying how they process the sensory inputs involved in steering their flight. A further scientific challenge is perhaps the most significant; how the sensory-to-motor transformation is influenced by motivation, and by internal states in general. Species can differ in their response to the same stimulus, as can animals of the same species trained to respond to the stimulus in different ways. How this is encoded in terms of brain activity is completely unknown, says Professor Heinze: "This will be the first time we can pinpoint for any species how intention is encoded in the brain, and that's why it's exciting".

This provides the impetus to share knowledge across species, and for the ERC project BrainInBrain to combine all the available knowledge about the central complex in insects. The *Insect Brain Database* is the vehicle for that, and its success so far has relied on the long-established collaborative culture of the relatively small field of insect neurophysiology. The challenges faced are all about building the database and its data management and publication functionality towards a self-sustaining community resource, integrated with other infrastructures for this purpose.

- *How do you intend to sustain the database as a community resource in the future?*

The ERC Starting grant is helping sustain the *Insect Brain Database* for the project duration. In the longer term, the biggest challenge will be maintenance costs. Database curation relies on voluntary efforts from participating laboratories. There are also costs of software development, and storage on Amazon Web Services that offers scalability and flexibility, both desirable for publishing large volumes of data online.

[9] Heisenberg, M. and Kaiser, K. (1995). The Flybrain Project. *Trends Neurosci*. Nov;18(11):481-3.

[10] http://www.virtualflybrain.org/site/vfb_site/home.htm

Heinze has a number of long-term sustainability strategies. The scientific use case is the most important, i.e. its effectiveness in enabling cross-species research on the central complex's role in processing intentional behaviour in insects. Heinze is not relying solely on prospective research funding, however. Other business models are being considered to help make a case for sustaining the curated database as a service. These rest on its value in generating further impact for the research.

3.4. Impact of open science practices

In its first year of operation, the *Insect Brain Database* has demonstrated the potential to make a more effective use of public research funding. "Now we are preserving all the raw data we have, to allow re-analysis, which will be of real benefit to the field", says Professor Heinze. Those benefits will make cross-species research on insect brains easier, its graphical searching and cross-species indexing will save researchers' time in locating published knowledge. The database is also contributing value to the published scholarly record, improving reproducibility by linking articles to underlying data, and enhancing the prospects for reuse by rendering this data in a form that helps researchers make further conceptual links, via the identifiers and standard nomenclature. The latter is also helping integrate the *Insect Brain Database* with world-level repositories including Virtual Fly Brain.

The 'one curator per species' principle ensures that appropriate expertise is involved in the quality control of additions to the database. It has also helped build momentum: "It's taking off now, there are around 30 people actively contributing data", says Heinze. Early contributors have been drawn from his personal network, but the curators are not a closed club. Any registered user can be a curator by adding data for a new species. They may also propose themselves as co-curators on the basis of their expertise.

Heinze foresees a more rapid growth in the volume and utility of the database, through several potential routes. More dedicated funding would support the contribution and curation of historically published data that may be reusable in other contexts. Another route towards a more rapid growth is to capture content closer to its source. Charging laboratories a subscription to use new functions, earlier in the research cycle, would offer a potential source of new content and revenue. These stand-alone applications would offer a closed and configurable environment researchers feel safe in. Linked to the *Insect Brain Database*, compatible with all data already present, the environment would be ready to be made available to the public part of the database with one click.

Professor Heinze also has plans to develop and market educational applications for schools; presenting interactive visualisations of the database content for a younger audience. This may generate revenue to sustain the resource and should also generate enthusiasm for the field among the potential researchers of the future.