European Commission

Horizon 2020
European Union funding
for Research & Innovation

Case studies on Open Science in the context of ERC projects – Set 4

December 2018

This document presents the fourth 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
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.1848198

LEGAL NOTICE

TABLE OF CONTENTS

SUMMARIES

Twelve Labours of Image Processing (Twelve Labours)

A commitment of one research group at ENS Paris-Saclay to document their research process systematically and publish research outputs that are easily reproducible gave way to a new open access journal, 'Image Processing On Line'. It is the first journal in the field openly hosting methods, codes and providing a possibility to conduct experiments and process images online. Created by Professor Jean-Michel Morel and his team during the Twelve Labours project, the journal is gaining its place in the computer vision and image processing fields. So far, it has not only proved to be useful in assuring further funding for the research group but has also been used widely by academics from various disciplines as well as other practitioners, from both public and private sectors.

Predicting Environment-Specific Biotransformation of Chemical Contaminants (PROduCTS)

Although chemical contaminants can pose significant risks to the natural world, microbes can reduce such risk through biotransformation. The PROduCTS project, led by Professor Kathrin Fenner at the Swiss Federal Institute of Aquatic Science and Technology (Eawag), is investigating, measuring and aiming to predict the outcomes of the biotransformation process through which microbes break down contaminants into different, typically less harmful chemical structures. Her team has established enviPath, a free online tool, available openly to everyone, to document the sparsely scattered valuable data about chemical contaminants, their transformation pathways, kinetics and end products. This is the first open database for such data in the field of analytical and environmental chemistry, and it is proving to be a valuable resource for the scientific community, public authorities as well as industry.

Layered Functional Materials – Beyond 'Graphene' (BEGMAT)

Michael Bojdys is an Associate Professor and Group Leader of the Functional Nanomaterials laboratory at Berlin's Humboldt University. As PI of the BEGMAT project, in the highly competitive nanomaterials field, he has led his group in a gradual shift towards data management and sharing; at first internally and now increasingly publicly. Spurred by the ERC's support for open science and his experiences as a postdoc working in the UK, his lab benefits from project efficiency and quality control by taking a formalised approach to research data management. Institutional infrastructure has been a key enabler and, in turn, the lab's RDM processes make it easier to share data to aid the reproducibility of published articles, extending existing practices of sharing supplementary material.

1. Twelve Labours of Image Processing (Twelve Labours)

Summary

A commitment of one research group at ENS Paris-Saclay to document their research process systematically and publish research outputs that are easily reproducible gave way to a new open access journal, 'Image Processing On Line'. It is the first journal in the field openly hosting methods, codes and providing a possibility to conduct experiments and process images online. Created by Professor Jean-Michel Morel and his team during the Twelve Labours project, the journal is gaining its place in the computer vision and image processing fields. So far, it has not only proved to be useful in assuring further funding for the research group but has also been used widely by academics from various disciplines as well as other practitioners, from both public and private sectors.

1.1. Introduction

The project Twelve Labours of Image Processing (Twelve Labours) set out to study a number of specific problems related to the methods of image processing. Funded by an ERC Advanced grant, this project was led by Professor Jean-Michel Morel at Ecole Normale Supérieure Paris-Saclay (previously Ecole Normale Supérieure de Cachan), Centre de Mathématiques et Leurs Applications (CMLA). During their explorations, Professor Morel and his team soon realised that they needed a dedicated platform to publish their research results, which comprised publications, code and online demonstration. Hence, the journal 'Image Processing On Line' (IPOL)[1] was established. It was the first journal openly hosting methods, code and providing the possibility to conduct experiments and process images online.[2]

- *How is IPOL different from other journals in the computer vision and image processing field?*

IPOL publishes relevant image processing and image analysis algorithms emphasising the role of mathematics as a source for algorithm design. It appears that it is the first journal in its field with extensive requirements for a publication. An IPOL publication is as precise and comprehensive as possible and includes several parts:

- a manuscript containing a detailed description of the published algorithm (method);
- an online demo, where the algorithm can be tested on data sets uploaded by users;
- a software implementation of the algorithm in C, C++ or Matlab;
- an archive containing online experiments.

Professor Morel explains the components of an IPOL publication in more detail: "A paper in IPOL is a triptych: First, there is the article that describes the method or algorithm that is being published. In

[1] http://www.ipol.im

[2] Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the articles in this journal without asking prior permission from the publisher or the authors, in accordance with the BOAI definition of open access. The articles are distributed under a Creative Commons CC-BY-NC-SA license: users are free to copy, distribute, transmit and adapt the articles for non-commercial purposes if they attribute the work to the authors and maintain this license. The implementations are distributed under a free software license; please refer to each implementation for more detail about the licensing option. The open access, copyright and license policy of IPOL is available here: https://www.ipol.im/meta/copyright/

principle, this is enough for anybody to reproduce the research. Secondly, it has an online demo, which allows users to try out the algorithms directly on their own images. Finally, there is the programme code that is also made available."

All the material is peer-reviewed and verified. This means that the reviewers of IPOL have significantly more work compared to reviewers of other journals. Their job involves checking all three parts of the publication and making sure that they are reproducible and coherent. Specifically, they are asked to check that the algorithms are reproducible in three different ways:

- by executing them remotely on the journal's server;
- by downloading the public code and using it;
- by checking that it can be reprogrammed easily following the detailed instructions (pseudocode) given in the paper.

The review process is organised by the editorial board comprised of 40 members. The editor-in-chief appoints the editor for a specific publication who then selects the reviewers. The editors and reviewers that are chosen should have dual expertise: in mathematics and in programming. The editorial board has evolved and expanded by inviting authors of IPOL publications to become members of the board.

In addition, there is a fourth part to a publication in IPOL, an archive. "This is a proof, so to say, that the publication and code are actually useful", notes Morel. The archive contains data of all online experiments carried out by the readers of the journal. When users go on IPOL, upload the images and try out the algorithms, these are then stored, if the user agrees[3], in the archive visible to everyone. Each online paper at IPOL has its own experimental public archive, where all experiments made remotely by anonymous readers are stored. Currently (July 2018), the archive contains some 260,000 experiments. The number of online executions is more than four times higher.

The journal also encourages a more informal exchange of executable algorithms between researchers by hosting 'workshops'[4]. Workshops are temporary publications of online algorithms that do not need to follow the publication rules but allow several research teams working on a project to interact directly. This avoids the 'software curse', which hinders research teams using different platforms or programming languages to communicate their algorithms easily between them.

Although Professor Morel started the journal more as an experiment, since 2011 it has evolved from a web platform to a permanent journal with 145 articles published and a further 21 submitted and 38 in preparation. He and his team were not sure whether the journal would generate interest among researchers, but it is now a fully financially sustainable journal.

1.2. Successful open science practices used in the project

- *What was the main goal of the project and how does it relate to IPOL?*

The goal of the project was to develop algorithms that would provide an opportunity for anyone to process images by running those algorithms and getting a result. "We wanted to establish a proof-of-concept that image processing could be an autonomous process and that such image processing

[3] As some images might contain sensitive information (images of persons and human body parts, scans of personal documents (passports), images of industrial processes, and more generally images for which the recipient has no authorization of diffusion from the owner), individuals may not wish to publicise them.

[4] http://dev.ipol.im/ws/

methods could therefore be published openly online", says Professor Morel. During the work in the Twelve Labours project, he and his team soon understood that such an approach requires a serious commitment and time resources from researchers – comparatively more investment than a usual journal paper since authors have to prepare the article, a fully executable code and an online demo. In addition, such an approach called for a thorough peer review process to ensure all the parts of the publication (the method description, the code and the demonstration) are scientifically sound. They realised that making the published methods into an official publication could incentivise researchers to engage in such an effort-intensive process. "IPOL was not our primary goal but it became an outcome. This way we could formalise the process and ensure that all the algorithms we put online would be verified by reviewers", says Morel.

- *Why was it important to you to have image processing methods published online and openly available to everyone?*

First and foremost, Professor Morel wanted to make image processing available to the public: "This is because everybody uses images and we wanted to go to a public place and not stay in niches of groups of researchers. With all images having become digital 30 years ago, people have access to computers and digital images. All cameras are digital, and everybody has a stake in and is concerned with image processing, including public entities, and private companies." If algorithms are clearly defined as well as easily accessible to all, they can be used immediately. This is in the best interest of researchers as well: "If you find a method, like this line segment detector method[5] that is very popular on IPOL, anybody can take it and use it for purposes that you did not even realise yourself existed." Hence, according to Morel, the best thing to do with such methods is to make them available to the public and have them used and publicised to the greatest extent possible.

Another motivation for Professor Morel and his group to make results openly available is research reproducibility. He describes his disappointment when as a group leader he had observed years of research work being lost due to irreproducible results: "When I did pure mathematics, I saw that if you produce a nice mathematical proof, but you write it in an obscure way, you will find somebody republishing the same results years later independently". This is what he and his team have observed in image processing as well. If research results are described clearly, outlining the entire process from the beginning up until the end and including clean and reproducible code, then it is easy for other research groups to take up the work and build upon it. Otherwise, people working on the same subject have to start from scratch, making the same mistakes and reaching the same conclusions. For Morel this is a complete loss of research resources. Therefore, he wanted to commit to doing reproducible research with his team: "I saw this as an opportunity to practice reproducible research which would be a big boost to the efficiency of my research group. Even if our initial idea of the public website [currently, IPOL] had not worked out, I thought that such an approach would add to the competitiveness of our group years later. People will be able to reuse the same ideas, algorithms and code, which are well documented and easily available." They now can reap the fruits of such efforts – they achieve more results by building upon research that was already done in the past and serves as building block for further explorations.

[5] http://demo.ipol.im/demo/gjmr_line_segment_detector/

1.3. Challenges faced and success achieved

- *What challenges are you facing in running IPOL?*

Although the overall number of publications in IPOL is increasing, the number of submissions has not been as high as Professor Morel and his team had expected: "We expected to reach a figure of 50 submissions per year initially. We aim at reaching 30 next year". They see a straightforward explanation for that. According to Morel, it is rare to have fully reproducible research results published in an article in the computer vision or image processing fields. Such papers comprise about 1% of all published articles in these fields. As in many other fields, the aim of researchers is to publish as soon as some results become available. This often means that results are published before the algorithms are sufficiently mature. The aim of IPOL is different and authors are asked to work out the code fully and make it easy for others to use and reuse. This is a time consuming and resource-intensive task. Authors submitting to IPOL are not always aware of the effort this will require. "We have worked a lot to make the editing part of the paper, code and demo easier. Our team included brilliant computer engineers for that purpose: Nicolas Limare and then Miguel Colom designed the online facility, now operating faultlessly some 200 different algorithms. If the algorithms and code are ready, then very little effort will be needed to put them online. We hope that this will yield more submissions and publications", says Morel.

Another issue that limits the number of article submissions to IPOL is the impact factor. Currently the (unofficial) impact factor of IPOL is 4.1. "This is good, but IPOL is not the top journal among the computer vision journals, and some people prefer to cite articles coming from journals of higher impact factor", explains Professor Morel. He further adds that this is a chicken and egg situation: "At some point we shall need an official impact factor. To make it official, we need at least 30 papers published per year. Currently we have about 20 per year. So this is a process and we should get there in 1 or 2 years". Initially, his group considered creating IPOL as a side journal to a well-established journal of a learned society. This would have helped by giving credibility to IPOL. However, the journal would have reached a smaller audience: "Only people from the applied mathematics community, or from universities that are paying for the journals of the societies would have had access to the online facility", notes Morel. This was incompatible with the mission to make image processing publicly available to everyone. Hence, Morel and his group chose the more difficult path of starting anew and of earning credibility and prestige through the efficiency of making all the results available to everyone.

1.4. Impact of open science practices

- *What was the impact of IPOL on you and your research group?*

Twelve Labours and the creation of IPOL were a success for the research team. After the project ended they managed to ensure a good flow of research contracts both from industry and public funding. Since they are a research lab, a certain ratio between public and private funding has to be maintained, and Professor Morel notes that at the moment this ratio is around half and half. In addition, public grants are sometimes linked to industry, for example funding from the French Space Agency with whom they collaborate on pre-industrial research. No matter whether public or private contract, Morel's lab always includes the condition in the research contract that all results have to be published online. According to Morel, this attracts further contracts, even from private companies, as they see that they will gain easy access to many reliable experiments online.

- *As the publications including code and demos are already openly available online, what motivates industry to still approach you for cooperation?*

Professor Morel explains in detail the motivation of companies to collaborate with his lab. First, the code published in IPOL is generally owned by the universities, research centres or researchers. Depending on the agreement authors get from their institutions, they licence their publications under a licence which is open for academic use, but in some cases also for industrial applications. Since the code and algorithms are published in a fully reproducible way, a company can buy a licence for the code from the authors or the institution (which is usually cheap) or they can reproduce it without permission, as it is openly available to everyone. This happens, and companies do use the published codes without permission. However, those who really want to use specific code for a serious industrial case contact the research team and ask for permission and collaboration. For Morel this is a sign that companies are interested in his lab's work: "When companies contact us we know that it is not exactly for the code, since it is openly available; it is for the knowledge we have in our research group. The publications from IPOL give a strong guarantee to the counterparts that they would have access to good researchers able to produce reusable code and usable algorithms. They also want more work to be done, including extensions to the code, or they want some new things from people who already produced good results and are likely to complete other tasks." Publishing one's research results in IPOL, therefore, is a sort of advertisement to research groups.

- *What other impacts of IPOL have you noticed?*

Professor Morel has observed that once authors publish a paper in IPOL and see the immediate impact, they return with another article and try to publish in this journal again. They have noticed that articles and other outputs are read and downloaded much more than in the case of other venues, and this in turn gives researchers more impact. In addition, publications are reaching a wider academic public than just from the image processing or computer vision fields, as well as public outside academia. According to Morel, IPOL publications reach doctors, surgeons and other practitioners: "People are hungry for seeing image processing methods at work; we have seen over the years that many researchers from private or public labs investigate the opportunities available through IPOL to see what they can use in their work."

With the help of IPOL, image processing know-how was expanded to a much broader set of academics and to many other sectors as well. Some examples include the police, army and the Office of Naval Research in the USA. The latter have been supporting IPOL since its inception. In addition, the journal is being used across the entire globe: "We made some statistics about which countries are using it the most. And with one exception, its use correlates to the gross domestic product of the country. The richer the countries, the more users, whether companies or academics. The exception is Russia, where IPOL is used much more, and I think this is understandable as the academic level in Russia is disproportionate to the GDP of the country", notes Morel.

- *In your view, what made IPOL so appealing to other disciplines?*

"Just because it is online and available to everyone", says Professor Morel. As many professionals need image processing, they search for available tools. One possibility is to buy software, which requires additional steps of downloading the program, the code and finding somebody able to programme in it. This is a complex process. IPOL provides an easier option. Everyone can go online, download a paper, upload an image and conduct an experiment on it. This intuitive process provides a direct window to people with no programming background, like medical doctors, to process images.

2. Predicting Environment-Specific Biotransformation of Chemical Contaminants (PROduCTS)

Summary

Although chemical contaminants can pose significant risks to the natural world, microbes can reduce such risk through biotransformation. The PROduCTS project, led by Professor Kathrin Fenner at the Swiss Federal Institute of Aquatic Science and Technology (Eawag), is investigating, measuring and aiming to predict the outcomes of the biotransformation process through which microbes break down contaminants into different, typically less harmful chemical structures. Her team has established enviPath, a free online tool, available openly to everyone, to document the sparsely scattered valuable data about chemical contaminants, their transformation pathways, kinetics and end products. This is the first open database for such data in the field of analytical and environmental chemistry, and it is proving to be a valuable resource for the scientific community, public authorities as well as industry.

2.1. Introduction

Waste chemicals released into the environment are detrimental to both humans and the natural world. Fortunately, microbial communities in the environment can reduce the problem by breaking down the chemicals – a process called biotransformation. Little is known about this process and in particular about what potential it holds to transform chemicals released to the environment. Professor Kathrin Fenner, a senior scientist and research group leader at the Department of Environmental Chemistry at the Swiss Federal Institute of Aquatic Science and Technology (Eawag), and her team are studying the timescale and the products resulting from biotransformation for different chemicals, locations and microbial populations. Through the ERC Consolidator grant Predicting environment-specific biotransformation of chemical contaminants (PROduCTS), they aim to derive predictive methods (i.e. model algorithms) from these observations and publish them online in a publicly-accessible database and prediction system. This project is highly interdisciplinary and combines the most recent technological and scientific advances in the fields of analytical chemistry, molecular biology and chemo-/bioinformatics. The final results of the project will have major impacts on chemical risk assessments, on the recovery of contaminated sites, and on the development of green alternatives for chemical contaminants.

To achieve its aims, the PROduCTS project conducts work along two main lines of research:

- Experimental branch, in which the team carries out biotransformation experiments with various microbial communities and different contaminants. They also characterise the microbes using state-of-the-art sequencing techniques in order to grasp a possible variability in contaminant transformation due to microbial community features. This provides a better understanding of why biotransformation proceeds readily in some instances in specific microbial communities and not in others.
- Data mining branch, in which the team aims to compile the existing biotransformation data on how fast and through which enzymatic pathways the chemical contaminants degrade. Researchers then use data mining techniques (machine learning methods) to draw insights from the compiled data.

2.2. Successful open science practices used in the project

When working as a postdoc at the University of Minnesota, Professor Fenner realised that a lot of information on environmental contaminants and their biotransformation was available across various registries and the scientific literature. However, such data were scattered in different documents and were not systematised. A user-friendly tool openly available to everyone that would put all these data together was missing. Hence, Fenner started working on creating enviPath[6], the environmental contaminant biotransformation pathway resource.

- *What is enviPath and how did it start?*

enviPath is a database and prediction system for the microbial biotransformation of organic environmental contaminants. It provides the possibility to store and view experimentally observed biotransformation pathways. The pathway prediction system provides different reasoning models to predict likely biotransformation pathways and products. The database was started by Professor Linda Ellis and Professor Larry Weckard at the University of Minnesota, with whom Professor Fenner worked during her postdoc. However, the initial version of the database had a major flaw: "There was a problem with the system as it was old and entering data into it was very hard. Only trained students who had previously worked closely with Professor Ellis could enter the data." Hence, Fenner started collaborating with researchers at the University of Mainz in Germany to develop a new kind of database. The key aim was to come up with a system that would allow collaboration and smooth entering of data by all research groups. enviPath gives such an opportunity and provides a user-friendly solution to research teams who want to submit structures and pathways into the database.

Currently enviPath has two main data packages. The first one was inherited from the University of Minnesota (handed over along with the rights to the database) and the other one is newly compiled by Fenner and her colleagues. "They both contain similar data (biotransformation pathways) and are pretty much the same in size", she says. "Essentially, these data are connected graphs describing how a chemical is transformed by microbial communities. They are maps of structures of how a parent compound transforms into different transformation products." The research team has also added information on the half-lifes (the speed of the degradation of chemicals) into the new package, annotating each one of them with the metadata that specifies the exact experimental conditions under which the half-lifes were measured.

The new data is mostly not based on the experiments Fenner's team carries out but on data coming from various registries. This is because, as she explains, "There is a lot of data that companies, producing various chemicals, have to submit to authorities before putting their products on the market. These dossiers, which are publicly available from the European authorities, contain information that is valuable for the database." Their work has therefore been focused on extracting the relevant data from text files (mostly PDF documents) and putting them into their database format, to make them automatically readable and interoperable by machine learning techniques. In addition, they are currently also using enviPath to store their own experimental data on contaminant biotransformation in activated sludge systems. This data package is still under construction and therefore not yet publicly available.

- *enviPath is clearly a significant achievement concerning the promotion of Open Science practices in your research field. Does your team also engage in other Open Science practices?*

[6] https://envipath.org/

Professor Fenner says that openness and transparency are important concepts to her and her research team. Alongside the work of building the enviPath database, they aim to publish their articles in open access, make the code open and share their research data whenever they can and whenever this is feasible. "Not all of our articles are openly available, because it is very expensive. There are a few journals that we favour for some lines of work and we try to publish in them as they allow open access", she notes. Usually, they choose the green open access route and make their publications openly available after 6 to 12 months of embargo. Fenner is aware that she could use the ERC grant for the gold open access route as well (pay article processing charges and have an article published in immediate open access), but she prefers the green route that allows her to make more money available for the team's research activities.

Regarding open sharing of research data, it is important to distinguish between the two kinds of data that the PROduCTS team works with. First, they collect the sequencing data on microbes; these data are made available when publishing an article. The sequencing data refers to the DNA and the RNA data that are collected through the biotransformation experiments in the project. Fenner explains: "In such experiments, we take samples of the microbial community and extract the DNA and the RNA (the genomic info about the community). After that, this information is read, and we get the sequencing data, which we make publicly available through dedicated repositories."

The other data, which the team is currently not sharing openly, comes from analytical chemistry measurements. These data are produced when studying how fast different contaminants degrade within the microbial communities. "We assess how fast that degradation happens by measuring concentration decreases over time, and we then analyse these data to get the kinetic rate constants. In these experiments we also use high-resolution mass spectrometry to help us determine into what kind of a product the contaminant is transformed", says Professor Fenner. "The raw data from these measurements is not helpful unless one has specific licenses or software which would allow actually doing something with the data. For these data, it makes more sense to share intermediate data such as rate constants and product structure spectra, and we are working on that by organising those data and creating metadata." Product spectra in particular are also submitted to MassBank[7], an open access database for high-resolution mass spectral data to which Fenner's department at Eawag strongly contributes. Such data organisation work, as she further explains, requires time resources. There are different people in the team carrying out the experiments. They have to pool all the results together into one database as well as ensure that the data is presented in the same format, and that the same quality criteria for data filtering are applied. They also have to log metadata describing actual experiments and where the raw data from these experiments are stored. Once all this is done, they would consider sharing the data openly. Fenner thinks that the recently established institutional data repository would be the most suitable option for storing and eventually sharing this kind of experimental and analytical chemistry data.

2.3. Challenges faced and success achieved

- *What difficulties did you face when starting enviPath?*

"Initially, there were many legal problems", admits Professor Fenner. In order to inherit the system from the University of Minnesota, her host institution (Eawag) and colleagues at the University of Mainz had to set up a license agreement. This proved to be particularly difficult as they are based in

[7] https://massbank.eu/MassBank/

different countries with different legal frameworks. “As a researcher you initially have zero understanding of any legal issues. Luckily, the host institution provides support as they have lawyers in place to handle these kinds of situations. Still, they are not used to handling such issues on a daily basis and it just took a long time to sort everything out”, says Fenner.

The other difficulty they experienced relates to the costs of the database establishment and maintenance. Fenner appreciated the fact that she had five years of continued ERC funding to do this kind of work, which, according to her, is hard to obtain from other funding sources. She also admits that maintaining enviPath will be a challenge after the ERC grant ends. Eawag is already partially supporting the database financially: “We do not have programming skills in-house, so we had to contract out some technical programming work, and the host institution made a financial contribution, paying 50% of those costs.” Fenner believes that in the future she will be able to cover the basic level maintenance costs from the Eawag contributions and her own professorship and funds, but she will need new grants for any major improvements or additions to the database.

enviPath is still in its early years and Professor Fenner admits that its use has not been as strong as she expected it to be. She thinks that this could be related to the data preparation that is needed before submitting them to the database. Researchers might not yet be ready to invest much time into such groundwork. “The number of people entering data has not picked up as much as we would have liked. Part of it relates to the fact that for the metadata you still need some adjustments. Depending on the type of system you study, you may need a slightly different metadata setting and permission. This has to reach us, and we can then tweak the database to allow for the different set up.” The team understands that it is still work in progress and that providing instructions to collaborators on how to prepare the data for the database is very important. “When it’s your own database, you are more motivated to do it; however, it’s not necessarily a highest priority for others”, says Fenner.

2.4. Impact of open science practices

- *What has motivated you to continue with an open approach?*

According to Professor Fenner the environmental chemistry field does not have a strong tradition of open access and open data compared to other areas, e.g. biological and medical research. “This is partly because the other fields are more ‘high profile’, people in those fields publish in more high-profile journals and there’s much more scrutiny to check whether the data was right and that it wasn’t fabricated”. However, she thinks that everyone should have access to publications and the underlying data. All published work should be available to everybody. Also, by looking across data, researchers can gain many more insights: “Having all these data within one database and having our own experimental data has had another effect. We started comparing biotransformation in different systems, e.g. how fast biotransformation is in activated sludge for different compounds versus in soil. By combining enviPath data from the literature with our experimental data, we started seeing many correlations between these two datasets, once you treat them right”, says Fenner. She adds that this brings completely new opportunities for predicting degradation in different environments, which they would not have discovered if they did not have those two datasets.

At the personal level, the addition of the new data package and making it publicly available online brought a lot of personal satisfaction to Fenner: “It was such a hidden treasure in those pdfs which we now make accessible to people.” Fenner is happy to see the data used in different contexts and the growing interest in enviPath evidenced by an increased frequency of emails and requests. Beyond that, she currently cannot point out any effects in terms of hard numbers e.g. increase in citations.

- *What has been the overall impact of enviPath?*

Despite all the obstacles, the enviPath database is receiving a lot of interest from various stakeholders. "We get many requests from people who want to hook-up their prediction tools or analytical workflows to our tool. Our database allows them to predict transformation pathways, therefore people are taking those predictions and integrate them into their tools", explains Professor Fenner. There is also interest from other labs and researchers who host databases and would like to download all the enviPath data and integrate them with their databases. The third kind of interest comes from industry, particularly from companies producing pesticides. This has pushed Fenner and her colleagues at the University of Mainz to establish a spin-off. Data on enviPath can only be used without having to seek permission for non-commercial purposes as they are published under a specific CC licence. The spin-off gives them an opportunity to licence out the data and enable commercial exploitation of them. Given the difficulties in gaining more funding for subsequent work, their hope is that having the spin-off company will enable them to receive money from the license fees for further development of the tool. Currently, the spin-off company is having first contract negotiations with a couple of commercial clients.

Work on enviPath also had some enabling impact at the national level. The national environmental authorities in Switzerland are concerned with the presence of pesticide transformation products in ground water and contacted Professor Fenner and her colleagues to start a project with them. "We took a long list of pesticide transformation products from enviPath and have been screening the ground water searching for the transformation products of those pesticides. Producing the list of suspect transformation product structures was one click, more or less, and saved the national environmental authorities a lot of time as it concerned thousands of transformation products. Otherwise they would have had somebody looking through the existing PDF documents for weeks and months to compile that information", explains Fenner.

Currently, the research team is also in contact with the US Environmental Protection Agency (US EPA). They are developing an environmental fate simulator[8] and would like to link it directly to the enviPath system. Fenner notes: "Of course, this is very complicated at the technical level as it concerns linking different software to each other and different database formats, but we hope to work it out eventually."

[8] https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=275872

3. Layered Functional Materials - Beyond 'Graphene' (BEGMAT)

Summary

Michael Bojdys is an Associate Professor and Group Leader of the Functional Nanomaterials laboratory at Berlin's Humboldt University. As PI of the BEGMAT project, in the highly competitive nanomaterials field, he has led his group in a gradual shift towards data management and sharing; at first internally and now increasingly publicly. Spurred by the ERC's support for open science and his experiences as a postdoc working in the UK, his lab benefits from project efficiency and quality control by taking a formalised approach to research data management. Institutional infrastructure has been a key enabler and, in turn, the lab's RDM processes make it easier to share data to aid the reproducibility of published articles, extending existing practices of sharing supplementary material.

3.1. Introduction

The discovery of graphene has broken new ground in the field of synthetic chemistry and nanomaterials over recent decades. This 2-dimensional form of carbon has unusual properties whose potential for industrial application earned it the reputation as a 'miracle material'. However, to be used in the fast-moving and competitive field of semi-conductor design, more basic research is needed to yield materials with the potential to compete with silicon, which is widely used in device manufacture. In the project Layered functional materials - beyond 'graphene' (BEGMAT), Professor Michael J. Bojdys seeks to develop a strategy for design, synthesis, and application of layered functional materials that will go beyond a small group of known materials in the graphene family.

The catalyst for this project was work by Bojdys in 2014 on a compound known as triazine-based graphitic carbon, highlighted as an emerging competitor for graphene. In 2016, his project BEGMAT started, which is an ERC Starting Grant project exploring the synthesis of compounds with the right characteristics for semi-conductor and other electronic applications. The work began at Charles University Prague, Czech Republic, and now continues at Humboldt University Berlin in Germany, where Bojdys is Associate Professor and Group Leader of the Functional Nanomaterials laboratory in the Department of Chemistry. The group has begun synthesising compounds that (like graphene) use widely abundant organic precursors and (unlike graphene) provide electronic properties when the materials are layered at the macroscopic scale.

3.2. Successful open science practices used in the project

Professor Bojdys is an enthusiastic supporter of the 'as open as possible, as closed as necessary' principle established for the European Commission's Horizon 2020 programme. As in the saying 'charity begins at home', for Bojdys data sharing begins in the laboratory, continues with the open access pre-print, and extends to depositing data in open repositories, where it can provide additional context to published work.

Research data is shared routinely at the group level, and often with academic and industrial collaborators. Data from the ERC project has been shared publicly for specific use cases, such as providing more context for a research article. The field of synthetic chemistry is highly competitive and there is strong commercial interest in the results. Open sharing practices have yet to become the norm. Nevertheless, pre-prints from Bojdys' lab are shared with supporting data on open repositories including Zenodo and ChemRxiv.

- *What motivates you to put efforts into good data management, and to share the data and your publications more openly?*

For Professor Bojdys, the adoption of more open practices is not just about compliance. The main driver for data management is to facilitate better control of data processing, and to support collaboration between project members. In particular, the effort spent to formalise data management for internal sharing within the project using a data repository has helped assure the quality of the data collection and analysis, and the PI's oversight of progress on the projects he leads.

The benefit of a data repository became clear to Bojdys through his experience as a post-doctoral researcher in the UK, where in his experience universities are further ahead than most in terms of their data management infrastructure: "At the University of Liverpool we had a Centre for Materials Discovery, shared with Unilever, and they had an industry-style data deposition framework, which meant that data collected on instruments like mass spectrometers would end up on central servers within this Centre automatically". This gave the opportunity for PIs to have a more comprehensive overview of the workflows to analyse the data.

- *How did you start using more open repositories to share data and supplementary material?*

Driven primarily by the ERC's guidance to deposit research data in open repositories, the BEGMAT team has begun to use the Zenodo repository to deposit primary data that would otherwise be handed over to journals and publishers as supplementary information. According to Professor Bojdys, this began with what was initially seen as a one-off experiment: "We had an article in Nature Reviews Materials and wanted to give further information about the verification of the composition of a certain material discussed in the review". Rather than handing over the information to the journal publisher, Bojdys deposited this information in Zenodo and linked to it from the article.[9] This was also a good example of data reuse, as the information consisted of data analysis he had conducted some years previously. In this case it provided "additional context for the readers to reassure themselves" about the material discussed in his article.

3.3. Challenges faced and success achieved

A key challenge for the BEGMAT project has been the necessity to mimic as far as possible the features of industrial scale data management platforms, using free cloud-based tools to compensate for the lack of institutional data management services at Charles University Prague, the initial host institution for the BEGMAT project, which has now moved to Humboldt University Berlin.

Professor Bojdys places high value on the host institution's support for the kind of infrastructure he had previously experienced at University of Liverpool. "The level of investment in setting that up and keeping it going easily surpasses the value of any individual ERC grant", he says. With the benefit of his experience in the UK, Bojdys sought to replicate as much as possible of the data infrastructure he had experienced in Liverpool at the first host institution for BEGMAT, which lacked the same level of investment in this area. He established the workflows to collect and process raw data using a combination of generic cloud-based platforms, Dropbox and Google Drive for working storage of raw data, Slack for team collaboration and notification, and Zenodo as a repository for final results.

[9] https://doi.org/10.5281/zenodo.1119088

"Well-thought out laboratory workflows that rigorously enforce practices for capturing, naming and organising data in a central storage facility are essential'', Bojdys says. Cloud services can provide the storage but, without automation, "the workflows can slow things down a little and introduce human error". Nevertheless, sharing the data within the lab increases the likelihood that errors will be spotted, and enables the PI to keep track of progress in the project.

The research team are also aided in their data management by electronic lab notebooks. "This is ultimately up to team members, but we have CambridgeSoft E-notebook and people do use it." Bojdys continues: "As PI, I request access to this, get a print out when a student leaves, and make sure that records are kept to the same standards as for a written lab-book". The digital form offers benefits in terms of indexing and searching. As PI he also encourages people to deposit their project-related measurements in one project file in the cloud storage, so there is context for the other files.

As a general rule, when sending out material samples to an external collaborator Bojdys shares as much information about them as is necessary for the project. Access for industrial collaborators is on similar terms as for academic colleagues, he says: "It is fairly free, but still curated." He admits doubting about the advantage of allowing third-party access to unpublished data, for the fear of getting scooped, and cites examples of a competing patent claim that would have been pursued by a commercial company, had the supporting data been freely available in advance of his own claim. He is, however, happy to share data as supporting information for pre-prints. From his perspective, the main barrier to maintaining an open data archive is the time that it would take him to maintain a hierarchy of access to data. That level of policing and managing of access, depending, for example, on individual roles and whether data has been published or not, is a service that he would look to the institution to provide.

For Professor Bojdys, the need to prepare sufficient description to understand the context of a dataset is an obstacle to sharing, except when the data underpins a publication. In his view the best form of contextual description for a research dataset is the research article that uses it as evidence. A potential disadvantage of this, however, is that negative or null results tend not to get reported in journal articles: "They usually describe the data that worked out well". Bojdys has addressed this issue by using supplementary material to discuss data that does not fit the story in the paper. He has taken this a step to further openness through the recent foray into Zenodo. As Bojdys puts it: "Describing in context the stuff that didn't work helps to convince the reader that all angles have been checked."

- *Apart from institutional factors, have you experienced any other obstacles to working openly?*

The rule that ERC projects funded under Horizon 2020 must provide open access to peer-reviewed publications within 6 months of publication has been problematic for Professor Bojdys' team. The main difficulty is with publishers, who commonly enforce a 12-month embargo periods for 'green' open access, and levy charges for 'gold' open access that he finds exorbitant. Bojdys also has challenged on his personal blog the practices of predatory publishers "who sometimes successfully capitalise on the gullibility of academics and also the legal obligation to disseminate data."[10]

The push towards open access has also encouraged Professor Bojdys to make wider use of pre-prints, e.g. in the ChemRxiv repository, and to use social media to make his research publicly accessible, through his twitter account and blog. He has faced some resistance to pre-prints in his community. The

[10] Bojdys, M. Parasitic Publishing (Blog, 8 Sept. 2016). Available at: https://mjbojdys.wordpress.com/2016/09/08/parasitic-publishing/

Journal of the American Chemical Society (JACS) was his and his co-authors' first choice of outlet for results they recently achieved on a photocatalytic process to extract hydrogen from water.[11] Unfortunately their manuscript was turned down by JACS as, according to its policies, posting research results or a manuscript on ChemRxiv or any other public database counts as prior publication.[12]

3.4. Impact of open science practices

- *Have you found any impacts on others of your effort to set up your own data management infrastructure using cloud tools?*

Professor Bojdys did not believe there to be sufficient experience in data management solutions to drive changes in the infrastructure provided by Charles University, the original host institution for his project, but he sees an opportunity for young research group leaders to try something new. Established experimentalist professors, he believes, rarely look at raw data from their students, making it difficult to convey the tangible benefits of such a solution to them.

The tangible benefit of data management for Bojdys' group is the additional oversight it offers him as PI, as all the raw data is accessible to him. "Even someone trying very hard to cheat with their data would find it very difficult to get that past me", he says. This improves the reproducibility of the research process, at least within the team. And there have been advantages for industrial collaborators. For example, to prepare contextual information supporting a proposal for a patent "we could get hold of unpublished data fairly quickly and fairly reliably".

Professor Bojdys makes no claim to an exceptional level of open sharing of all research data, but believes that the moves towards openness, the instilling of rigorous data management practices in his group, and commitment to public communication of his research get noticed. "The additional workload is justified, as you are giving a signal that the research, the PI, and the group put an emphasis on high quality. This gets picked up".

- *How has your use of pre-prints affected dissemination of your research?*

Pre-print sharing has accelerated the sharing of results with present and former colleagues. A case in point was the ChemRxiv paper on photocatalytic hydrogen evolution mentioned earlier. Only two days after the pre-print was posted by Bojdys' team, his former colleagues in Liverpool were posting results on the same site to rival their own, claiming further advances on similar materials. This led to lively exchanges via Twitter.[13]

- *Have you noticed any benefit from open access in terms of citations of your work?*

Bojdys expects that open access to results may boost the article citations but acknowledges this would be difficult for PIs to measure. "I hope at least that people are not citing work they have not read, and I know from colleagues' comments at conferences that if they are working in industry they often don't

---

[11] Kochergin, Y. et al. (2018). Exploring the Goldilocks Zone of Semiconducting Polymer Photocatalysts via Donor Acceptor Interactions. Available at: https://chemrxiv.org/articles/Exploring_the_Goldilocks_Zone_of_Semiconducting_Polymer_Photocatalysts_via_Donor-Acceptor_Interactions/6210110/1

[12] JACS have recently announced that manuscripts that have been posted as preprints on ChemRxiv, bioRxiv or arXiv will now be considered for publication: https://twitter.com/J_A_C_S/status/1031300824889208833

[13] Michael J. Bojdys twitter feed. https://twitter.com/mjbojdys/status/992381926907482112

have access to journals. So in that respect it's a big boost. It's very difficult to quantify but for example if you tweet a link to an open access article you can see how often it gets picked up".

- *Have there been impacts from your other communication activities?*

Professor Bojdys has been active in organising outreach events, for both academic audiences and the general public. For academic audiences he launched FuNMat[14], a series of annual symposia bringing together researchers in the areas of materials design, nanomaterials, biomaterials, electronic and optical materials. They have worked well in bringing colleagues together and making new contacts for potential collaboration. For public audiences, Professor Bojdys organised a series of events at Charles University under the heading 'The Pint of Science in Prague', inspired by similar events in the UK. Captured on YouTube[15], the events sought to bring members of the public together with scientists to discuss their latest research. He believes the public outreach events have successfully engaged their audience: "The Pint of Science managed to attract lay interest from parents and relatives of students. They would come along to learn about science in a relaxed atmosphere". This brought tangible benefits in terms of student recruitment; "they could see we were doing interesting work" says Bojdys, "and in terms of turnout it was clear they were events that people enjoyed".

In the longer run, and with institutional support, Professor Bojdys sees a natural progression from group-wide data management to more open access to data: "The opening up of these treasure chests". With sorting and contextualisation, open access to nanomaterials data would enable researchers to match particular kinds of infra-red or x-ray data to the samples described in publications. Developing group-wide data management and finding the opportunity to discuss and advocate the tangible benefits "without sounding like you are lecturing on the topic", takes his group one step on the way towards more open science.

[14] https://web.natur.cuni.cz/orgchem/funanomat/index.php/events/funmat-2017/

[15] "A Field Guide to the Science-Pint". Available at: https://www.youtube.com/channel/UC8RezSZ3NhWrfAgGbeva37w