CASE STUDY: EPISTEMOLOGY OF DATA-INTENSIVE SCIENCE

STUDY ON OPEN ACCESS TO PUBLICATIONS AND RESEARCH DATA MANAGEMENT AND SHARING WITHIN ERC PROJECTS
1.1. Introduction

Sabina Leonelli is Professor of Philosophy and History of Science at the University of Exeter. Professor Leonelli is a keen advocate of open science. She is member of the Working Group on Open Access Publishing of the Philosophy of Science Association; and of the Open Science Policy Platform (OSPP), a body advising the European Commission on policy formulation and implementation in this area. Leonelli is also PI on the project Epistemology of Data-Intensive Science, which began in 2014 and extends to 2019 supported by an ERC Starting Grant. The project explores the philosophical assumptions underlying the choice and use of taxonomies, theories, models and experimental methods in data-centric biology and biomedicine. It also looks at how relevant data infrastructures, and their division of labour, are involved in the collective modes of enquiry and scientific modes of understanding in these fields. A further angle to the research considers how tools for data dissemination enable integration and discovery.

Three years into the project, its detailed empirical investigations are already providing crucial insights on the complex conditions under which data can be made "open", as required by current EU Open Science policies. This applies to the data managed in the project itself, as much as to the data the project is tracking on its journey through biological and biomedical infrastructures to wider dissemination and reuse. The team are making their observations available through open data repositories, and endeavouring to apply their insights about data governance to their own data management practices. Those insights include the role of security and ethical concerns in the strategies used to integrate biomedical data; the ways in which labels, models and visualisation tools used by databases affect the interpretation of research data and their use as evidence; and the implications of the various ways in which research data are being made 'open'. The project is framing a new epistemological perspective on scientific research, which places data at its center and helps to explain the impact and implications of data science and "big data" for contemporary research efforts across different disciplines.

1.2. Successful open science practices used in the project

- What has the research accomplished so far?

"We looked at lots of different cases of reuse of data, and how data infrastructures actually mediate that reuse", says Professor Leonelli. The Data Science project has charted the obstacles encountered by data as they travel from the lab to the outside world, via research networks and the tools, databases, and repositories that transform them on their route to reuse. Publications by Professor Leonelli and her team demonstrate the significance of situations in which these data are missing, absent or inaccessible for future research. The studies document how research communities are organised in order to take advantage of large datasets and related digital technologies. "For example in the biomedical part we focused on the SAIL Databank Secure Anonymised Information Linkage, which holds a lot of patient and clinical data since about 15 years and has real expertise in how to manage data. It is a very important case study for us on how reuse can be mediated both for ethical reasons and to make reuse more efficient. Another case is MEDMI which involves the University of Exeter, the MET Office and others in a data infrastructure that is a mash-up of medical, health and climate data to run correlations. A third one is about somatic mutation data in cancer genomics, through the EBI COSMIC database, and involving very different issues. And then we have the biological databases."

The research has produced many more case studies than orginally envisaged, and they have involved international collaboration on a broad scale. "We have worked with people in Africa, in the United States, Europe, and Australia throughout this project. Now this is reflected in a volume we are putting together where each chapter focuses on a particular realm of practice, including social sciences and many of the
natural sciences”. This edited volume will be the second book to emerge from the project, following Leonelli’s 2016 work ‘Data-Centric Biology: A Philosophical Study’. A third is in the pipeline, described as “an over-arching view of what those studies teach us philosophically about what has changed in the methods of science, in the role of science in society, in how we conceptualise science, now that all these big data infrastructures are being used on a massive scale”. OA publishers are being sought for both.

1.3. Challenges faced and success achieved

- *Have there been any aspects of data management in the project itself that have proved challenging?*

We spent a lot of the first year of the project setting up for using open data. There are big issues in open data for projects like this. When we talk to people about their difficulties in managing their data there are big methodological issues about following the data and picking up the materials you want. And we have all kinds of data ourselves; relating to the archives people have when they document what they have done in the past, and a lot of documentation online, including their conferences proceedings and publications. Of course we also have a lot of interview materials and documentation from visits to laboratories and other locations where researchers are based.

It is a very diverse set of data sources, which is very unwieldy to deal with particularly because this is very much in the tradition of qualitative research and ethnography, where a lot depends on the point of view of the researcher and on their interpretive judgements, so there is never really anything objective about the data. However, in this particular case it is very much dependent on the interaction of the researchers and the interviewees. It is very situated and something we have seen is a problem for others and of course for us. It is a really big issue and question of how to compile all this data to be open.

- *How did you approach these issues about making the interview material open?*

What we have tried to do is to at least provide our interviewees the option. We wanted them to think about whether they want their interview to be part of a corpus of open data, and give them the option to opt in to that. What we have done was elaborate, pilot, and then widely use a particular type of ethics consent form. This consent form (1) has an additional part describing briefly the forms of dissemination used in the project and ask the respondents whether they would opt in to making the data and the transcripts available online either anonymously or not. If they choose to anonymise transcripts that means a lot of references to individuals and locations get elided, or they can also choose to make the transcript available online and attributable to themselves. There are all sorts of options here and it was quite interesting to put them together. Interviewees are very happy to have the option that, to comply with Data Protection law and our own ethical requirements as researchers, any moment they express opinions about other individuals or institutions during the interview these can be elided.

Doing it this way means we have ended up with a corpus of open data for much of the case studies–about a half of the interviews. What typically happens is that the more interesting an interview is for our purposes, the less likely it can be made open. The simple reason is that the changes around data sharing are cultural,

---

1 The consent form is accessible from the project website, datastudies.eu, at: https://datastudies.eu/images/downloads/other/Ethics_consent_form_DATASCIENCE.pdf
involving tensions and conflicts between parts of academia and beyond academia, so there is some data we cannot make available openly without betraying confidentiality agreements with the interviewees.

I think it is a very interesting exercise, because a lot of people who do my kind of work think that these are such sensitive data that there is not even a question of making them open. I actually think one has to be very careful around making assumptions about what it means to have open data for qualitative studies, as it is actually possible. In a study like ours, a lot of the more technical interviews for example about people explaining in their own words how they make certain databases work for them. Such information may not be particularly controversial but it is actually extremely useful for other groups who are interested in understanding how open data works and about its history. This way you can generate a corpus of materials that can be beneficial to other. For example, others can use the material and do discourse analysis on it or analyse the user stories. It is worthwhile.

- **How will you make the case study data available?**

What we guarantee now is to make the data available on Figshare. There is a large corpus of data and a dataset relating to each study. We are now starting to publish and as we publish each paper a related dataset will be put online. There is not that much software that is well suited to this kind of social science data and we were not impressed with what’s around. We are planning to do it with Figshare as you can insert the metadata and it is not particularly restrictive.

Then, we will make lots of entries to make it easier to locate data in a particular case and particular area, but also make them part of a collection. Hence, if someone wants to see all the data relating to the study they can also see that. Curating the data to make it findable and accessible is extremely laborious. It requires going through the transcripts carefully to make sure that everything we are going to make open is ethically okay. Sometimes we have an additional layer where the interviewee has also asked to double check it first. It takes a long time so we are staggering release of the data to coincide with each paper. Of course when we put the grant together, open data was not quite as recognised by the European Commission, so the ERC did not have provisions to give us time to think about that. So any time that goes into this means less time to do the analysis and the research, which is the usual problem.

- **You have written extensively about issues in making data mobile and how some data can travel easier than others to be used in another context. Does the effort involved in representing your own case study contexts to make them understandable by others make your data less mobile?**

In the data I deal with this it absolutely obvious that data provenance is essential for any kind of re-analysis. The work I do is partly historical, and a lot of the relevant metadata will be in the published article, because that is the story of where the data has come from. So this makes it clear that data reuse needs to be aligned with reuse of the publications coming out of the project, because without that you will not be able to understand the context of data collection. This is not so dissimilar with what is happening in the natural sciences. What you get in some fields like economics is the feeling that you could reuse the data without much provenance information, while certainly in the biological sciences anyone who does very serious work with big data analysis is adamant that they’re not entirely sure what the data mean or how to manipulate them properly without being able to disaggregate the data and dig back to where it comes from. It’s important everywhere, it is just that in qualitative social science it is unavoidable. The use of numbers in other disciplines may hide the fact that you need to go behind the numbers to specific situations.
What is interesting, is to recognise there are kinds of data that can be portable, and other sources of information that may not take the form of official data but are crucial to your interpretation, and in qualitative research that’s impossible to take out, and it can get to things that measurement does not give you. It also means there’s a limit to how much of it can be made open. But the idea in open data is to focus on data sources that can be reused in other projects. They may contain interesting information about how certain resources were put together and what lay behind it. And this is the kind of information that can be very useful for other people thinking about these issues. Thinks like people explaining their motivations, and what were the challenges at that particular time in that particular location and in that field. This is the kind of thing that can be important to the history of curation in certain fields, and also the differences between fields, over time, is very important as we are in a very transformative moment in science, and it will be interesting to look back on how different researchers were articulating their position.

1.4. Impact of open science practices

- In some of your publications you highlight differences between making data accessible and actually reusable, and describe capabilities people need to reuse data. Does this work help the data infrastructures themselves in understanding what they need to do to help that?

Yes. Metadata will always be partial, the question is what rationale is given for them. This is the big difference between data dumped in an institutional repository without any possibility to search on it, and something like a database organised using the ISA-TAB metadata standard, because at least you know what you will be able to search for and the limitatons of it. Certainly among the best cases of reusable data we have looked at are where the people involved in curating know the data inside out – and can advise on how to formulate the search. In our kind of social science we are very much behind. We are stuck in thinking about simple archival labels – authors, year, things like that. I think it will be interesting to signpost these data using keywords to help reconsuct the analysis. The approach we use is very much a Grounded Theory one, where we are open to new things coming up. For example, security concerns and what they mean for data management. Hence, we did not want to be constrained by analysis software that requires fixed coding frames.

- That grounded theory approach of finding patterns in data to follow lines of enquiry iteratively, is it offering lessons for other disciplines that use big data?

I wish it were the case. The problem with some of the other disciplines that use big data is the expectation that the patterns they spot will be acquired automatically. Certainly people see things they have not seen before, in biology and other cases, because of the algorithms they have applied. But the second important step is to verify what that means and how it can be justified, and that requires very specific judgements from the investigator. And that tends to be forgotten in the rhetoric around big data, that this is a very important part of people working with large datasets. You need human judgement by experts in the subject to judge whether a pattern makes sense or is just an epiphenomenon of the instruments, or the sampling approach.

- How is that interpretive ability and context sensitivity emphasised in developing the research skills for data science?

It is absolutely not emphasised. I teach people about what it means to provide contextual information, and how that squares with the application of algorithms in the technical part of data science. There is a huge danger in data science going in the wrong direction, in selling it as a branch of engineering. Part of the reason I do policy work in relation to my project is precisely to avoid this conceptualising of data science as just a question of devising of cleverer algorithms. Anyone we interviewed who does interesting work with big data does not operate like this. Of course it is important to have better software, but the fact there has to be
judgement and contextuality involved in every stage of analysis is really important. A lot of the data staff training programmes we see have a danger of only dealing with the technical parts.

- **Does the ERC-funded work also feed into your work in the Open Science Policy Platform?**

Yes, I am in the OSPP representing the Global Young Academy, the viewpoint of established researchers not in the most senior stage of their career, and they very much have a stake in what is going right now. I am also working as an expert for the European Commission on a Mutual Learning Exercise developing a road map for open science in 14 countries, and the ERC project has very strongly informed that. As well as highlighting the issues mentioned above, the project has informed several reports by Professor Leonelli. They include a report on improving the incentives and rewards available to researchers for working openly; one of the key issues for the OSPP, and for the broader research community. A further report for the Swiss government describes the impact of big data on biomedical knowledge production.²

---

² Mutual Learning Exercise (MLE) on Open Science: Altmetrics and Rewards. Background/Challenge Paper No. 3 Incentives and Rewards to Engage in Open Science Activities