Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

AI3SD Interview with Dr Zosia Beckles
25/02/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

04/01/2022

Humans-of-AI3SD:Interview-14

AI3SD Interview with Dr Zosia Beckles
Humans-of-AI3SD:Interview-14
04/01/2022
DOI: 10.5258/SOTON/AI3SD0177
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Interview Details

Title	AI3SD Interview with Dr Zosia Beckles
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	ZB: Dr Zosia Beckles - Unviersity of Bristol
Interview Location	Online Interview
Dates	25/02/2021

2 Biography

Figure 1: Dr Zosia Beckles

Dr Zosia Beckles: 'Make a data management plan!'

Zosia Beckles is an experienced Information Professional with a background in health informatics and research data management in the higher education and research sectors. She currently works with Library Research Support at the University of Bristol, providing training and support to academics in research metrics/bibliometrics and sensitive research data management, including the development of a new dataset disclosure risk assessment service to enable safe publication of sensitive data.

In this Humans of AI3SD interview she discusses the potential offered by more open data sharing, the risks posed by reuse, the impact of Covid-19 and the importance of data management plans.

3 Interview

MP: What's been your path to where you are today?

ZB: I'm currently a Research Information Analyst at the University of Bristol, part of the Library Services team. We support researchers at the university in all areas of research but I specialise in researchers who are working with sensitive data. My first degree was in the sciences, and then I switched my focus. I knew I wanted to work in academia, in research, but not as a researcher myself so, initially, I started working in academic libraries. I got an MA in Library and Information Sciences from UCL and from there I started working for the Royal College of Obstetricians and Gynecologists in a team that produced clinical guidelines for NICE. That was really, really interesting to me because it's research, it's very practical work. You're looking at the evidence for specific medical or more qualitative interventions and it's all about making people's quality of life better. Being able to feel like I was directly contributing to something like that was incredible. That's where my interest in the research data side of things came in because, at the time, the amount of supporting data that was available, even in medical research, was much less and it was much more difficult to get hold of it to inform clinical reviews. That was the genesis of my interest in data management specifically, as a part of informatic support in general.

From there I started working for the University of Bristol, in a research data management specific team. In the last two years that has transitioned into a specific sensitive data support role. The awareness that people need support in that area has come much more to the forefront of people's thinking and the library wanted to have somebody who could specialise in that area and give researchers advice on working within GDPR, dealing with data ethically and safely, and making sure that they were thinking about these things as early as possible.

MP: How does your role play out on a day-to-day basis? Are you embedded in specific projects? Do people come to you and say, "Ok, I'm going to be working on this project and the data is sensitive. How do I handle it?"

ZB: We make ourselves visible so that people know to come to us when they're at the point of applying for grants, because most funders these days require a data management plan. They'll come to us at that point to get their plan reviewed and, if it involves sensitive data or human participants at all, it'll come to me.

We also tend to get involved at the end of a project, when people are coming to publish data. And that can be a bit of a problem because, obviously, there can be quite a gap between what people plan to do and what actually happens. If we are not aware of what that drift is, it can be difficult to support researchers to publish the data in the way that they want. They may not, for example, have all the consents and things like that in place. So towards the end of the process my job can be helping researchers to make the best they can of the situation with regards to data publication, given the constraints they're working within.

In terms of being embedded, I do have close relationships with certain groups. So, for example, my whole team has worked closely with the Avon Longitudinal Study of Parents and Children. And there's a big engineering study called SPHERE (Sensor Platform for HealthcarE in a Residential Environment), about embedding sensors within homes to enable people who might have issues with falling or something like that, to live in their own homes and be passively monitored so that an ambulance can be alerted if they fall, for example. We've worked quite closely with that project to make sure that their data is being managed effectively and we can get the most use out of it afterwards.

MP: What's the general level of knowledge that you're coming across and is it changing? Are people becoming more aware that they need to be concerned about data management?

ZB: It varies a lot from field to field. So, for example, in health sciences, researchers are usually pretty well informed about the need to share data and the things they need to do to make sure that they can. In other fields where that kind of open data, open research, drive has been a bit later in coming, there's understandably less awareness of those sorts of things. We're seeing more of an awareness of it in early career researchers, as the message seems to have percolated quite well in the postgraduate researcher community.

MP: What potential is offered by more open, ethical data sharing and data management?

ZB: Firstly, the ability to reuse research data, to maximise the value of data that may have been incredibly difficult, complex and expensive to collect in the first place, is really important. It's not just about maximising the funding, but also, where it's human data, maximising the value of that participant input. People are giving up their time and their sometimes quite sensitive information, to benefit themselves if it's a relevant study, but in general to benefit society as a whole. It's only right that we make the most of that where we can.

Being able to reuse that information can help to reduce the burden on research participants in future. People definitely get survey fatigue even when it's just shopping preferences or whatever, and that's got to also play a part when it comes to census collection and those sorts of big longitudinal studies. If we can make the most of the information that's collected once, then I think it helps to generate goodwill. And means that people are more willing to take part again, if they know the most is being made out of when they have contributed.

MP: What are the risks, particularly around, for example, sharing personal data, and anonymising data?

ZB: For me, this is all about balancing the need to maximise the value of that research data once it's been collected with protecting the privacy of the research participants. There are inherent risks in wanting to make data as open as possible. There's the issue of being able to re-identify participants just from what's in the dataset itself – but also from potential data linkages and things like that – and then linking multiple datasets together and being able to exploit vulnerabilities and identify people that way. Obviously, you've got the risk to participants in that. If they are re-identified, not only could it be harmful and damaging to the participant in question, but also it's a very bad look for research as a whole if taking part in something negatively affects a participant. It's going to be much harder to recruit people in future. The way to address those risks is not necessarily anonymisation; while it's very important, you can't ever anonymise something fully. Another way to protect participant confidentiality, while allowing researchers to make use of data, is through controlling the data environment. That might be through technical solutions, such as the SafePod network. Or it might be through data access agreements which limit what people can and can't do with data and set out terms for how it should be stored and dealt with.

Controlling the data environment is a much more practical way to address that inherent risk in data sharing than trying to go down the route of anonymising the data too much. I think you can try to fiddle with the data and reduce the risk in data itself too far, and take your eye off these other solutions that are available.

MP: How widely used are these tools and solutions that help manage the data environment?

ZB: With the launch of the SafePod network, I hope that secure areas for using data will be more available to more people. The idea of having data access agreements to govern controlled access to research data is hopefully becoming more widespread in the smaller repositories as well. Bristol's been at the forefront of that. We're one of the first institutional repositories in the UK to offer controlled data access and to have a data agreement over and beyond a standard license, which is much more dependent on the goodwill of the user. In contrast, these data access agreements are formal contracts and so there's much more heft behind them. The process of having to get an institutional signatory really helps to impress upon users the importance of taking care of the data and using it appropriately, in a way that if you can just download a file from a website at best, or check a box and say, ”Yes, I promise I'll do this” it has less impact. Which isn't to say people wouldn't follow those rules, but I think the process of going through a formal application to use data does help to clarify the seriousness of the care that has to be taken with data.

MP: How do you see it developing in the future? Eventually, do you hope that you will make yourself redundant and all of this would be embedded in researcher education? Or would it also be more automated?

ZB: I think in terms of researcher education, I would really like to see more attention to data management as a whole. And, specifically, issues around consent and ethical data collection more embedded into postgraduate training and education.

On a broader level, one thing that I would be really interested in seeing is getting participants more involved in how their data is reused. This is something that I picked up from working on NICE clinical guidelines where there's always at least one, usually two, patient representatives on the guideline committee, developing clinical guidelines. I think that's really important and it's something that I don't see in data reuse. For example, we have a data access committee that includes senior academics and people from IT services and research governance and ethics, all these groups. But we don't have anybody whose specific role it is to consider the interests of the research participant. I think it's understood that everybody should be doing that, but I do wonder whether there's value in having a specific person, much in the way that you'd have a lay rep on an ethics committee. Just to make sure that the people whose data we're considering are kept at the heart of those decisions about what's going to happen, particularly when you're talking about secondary uses that may be completely unrelated to the reasons for which the data was initially collected. Obviously, you want to be able to reuse data and make connections in ways that perhaps hadn't been thought of initially – and that can be incredibly valuable. But at the same time, it is important to consider whether the interests of the research participants are at the heart of that decision-making.

MP: More broadly, what surprises you about your work?

ZB: It's the breadth, the scope, of research that I get exposed to in the course of mostly dealing with incoming requests for data at our repository. Just the breadth of research that goes on and the uses that people find for data, when I would never have thought that these things could be applied in different ways. It's really inspiring and incredible to see this kind of research happening in real time in front of you.

MP: Has there been any impact on your work from Covid-19?

ZB: There has been a change in the number of requests that we're getting for datasets. As it has become more difficult to collect primary data, researchers are shifting to work with secondary data more. The other impact has mostly been shifts in the sort of thing that people are asking for help with. We're starting to have many more queries about working with and collecting data online and how to work with it safely and securely. And the technicalities of recording things on Zoom and where that's going to be stored and those sorts of things. It's been quite a steep learning curve in figuring out how all these tools work. Where they store their data, what servers they pass through in the EU, in the UK, what's the impact on various bits of legislation?

MP: If you could wave a magic wand and make sure that all researchers understood one key thing about managing and sharing data ethically, what would that be?

ZB: The importance of consent – or at least informing participants of exactly what your plans are for data sharing. That's the barrier that we come up against the most. The structure around informing participants about the nature of the study they're taking part in is quite well established. Research ethics committees are well aware of the things they need to look out for in terms of informed consent in that respect. But a lot of the time, researchers and research ethics committees don't necessarily think about what might happen with that data after the end of the study. So we can get to the stage where researchers would like to publish data but can't, or at least will have difficulty in doing it, simply because they never mentioned it to the participants, or worded the consent form in ways that it's likely that the participant understood that their data would be kept completely confidential. Maybe that wasn't entirely what was intended, but that's the wording that's in front of them. Researchers need to understand the importance of thinking about what they want to do with the data in the long term and making sure that their participants are informed.

MP: What advice would you offer to early career researchers?

ZB: Complete a data management plan! That is the one thing that will get you thinking about what you want to do and the steps you need to take to ensure that you can. From initial data collection through to storing your data securely, documenting it, and thinking about what you might want to do with it in terms of sharing – completing a data management plan would get you to think about all of those. And if you do it once, you've then at least thought about it. So you've got a vague roadmap of the steps that you need to take in order to get underway with your project and be able to meet the requirements that you've set – for example, if you've said you will publish your results in a particular journal that requires you to publish supporting data, you need to have that in your plan.