AI 4 Science Discovery Network+

AI4SD Interview with Professor Henry Rzepa
29/01/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

11/08/2022

Humans-of-AI4SD:Interview-30

AI4SD Interview with Professor Henry Rzepa
Humans-of-AI4SD:Interview-30
11/08/2022
DOI: 10.5258/SOTON/AI3SD0238
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 **1**

2 Biography **1**

3 Interview **2**

1 Interview Details

Title	AI4SD Interview with Professor Henry Rzepa
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	HR: Professor Henry Rzepa - Imperial College London
Interview Location	Online Interview
Dates	29/01/2021

2 Biography

Figure 1: Professor Henry Rzepa

Henry Rezpa: 'We don't have to be a profit centre for large multinational publishers'

Professor Henry Rzepa of Imperial College London started as a synthetic chemist in 1971 and then became a computational and information scientist and a spectroscopist. These research activities have generally generated large amounts of data and early on he became increasingly concerned that this vital research product was rarely treated as what is now called a first class scientific citizen. Since 2005 he has been trying to elevate his group's data to this status, using a combination of ELNs (electronic laboratory notebooks) closely coupled to what is now three generations of data repositories, to try to achieve its FAIRdom (Findable, Accessible, Interoperable and Re-usable).

In this Humans of AI4SD interview he discusses his life in data, the barriers to data sharing and the role of funders.

3 Interview

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

HR: In 1987 I was a regular mid-career academic, and I was invited onto the editorial board of the Royal Society of Chemistry. At the first meeting I attended, I mentioned this little area called data. We didn't know much about it but we suspected an electronic era was coming. At that point in time, whatever data was present in a scientific article was firmly part of it. In some infamous cases, tables and tables would be broken out over several pages, they were that big. We were beginning to recognise that more and more fields were generating copious amounts of tabular data, which would become a problem. Back then, when people had too much data to publish in the paper, they put it in an envelope and, if the article was accepted, it got packed off to the British Library at Boston Spa, where it was archived. We were looking to find better solutions than these basements of decaying paper.

In 1993, the web came along and I thought about the potential it might have for this brief. I wrote an article for the journal, in fact, highlighting the potential here. The referee said that the paper was very unusual, because it didn't contain any chemistry, but that it held some potential for chemistry's future. It ended up being one of the main things I'm proudest of doing in my career.

Between 1993 and around two or three years ago, Electronic Supporting Information (ESI) developed in a very ad-hoc manner. A lot of people latched on to Acrobat, and decided that it was going to be the permanent holder of this ESI, but it was never controlled, never discussed by a committee. Various informal rules developed around ESI, which were produced by the users on the basis of whatever was the shortest possible time in which the information could be produced. Then a few years later, PhD students started producing their theses and the rot truly set in. It reached its peak, to my knowledge, around three years ago when the first 1,000-page Acrobat file was submitted with a paper.

During this time, ESI was uncontrolled and few people gave it much thought. I was still publishing with the Royal Society of Chemistry, and I thought I had better demonstrate my best practices, so I started sending them stuff in a modern way. In 2006, we also started storing data in a data repository and, at that time, we were one of the few groups to have a data repository in chemistry. We also generated persistent identifiers, so all you have to do is quote the identifier and someone could have access to your data. Additionally, we created visual forms of data: not just static pictures but rotating molecules and interactive components. To date, I've produced about 80 of these, published in about 60 articles, which is much more preferable to a 1,000-page Acrobat file.

These visual forms of data are not simply representations, but quite a sophisticated toolkit. There's zillions of things you can do with it. That's the I of the FAIR principles: interoperability. Accessing data doesn't just mean getting hold of it in some inscrutable and unreadable file format, but accessing it in a form that you understand. Our approach anticipated the FAIR principles before they were formally developed.

MP: Why do you think this approach to data has not been more widespread?

HR: Two years ago, the FAIR community in chemistry, which probably consisted of around 15 to 20 people worldwide, were organised by the National Science Foundation in America to get together. We met in a hotel in Orlando, Florida, and over two days we discussed how we can

make people more aware of it. How can we persuade more people to start using it?

One of the people present was an editor from the journal Organic Chemistry. As part of his contribution, he ran a pilot project on the journal in which people submitting papers would be asked to click on a link that initiated a FAIR data upload. With around 1,500 paper submissions to that journal every year, only around 100 clicked on the link. The link was very ambitious: it basically asked the submitter to share a folder of data, which was then zipped. This journal now has 100 zip files as part of that project. So Organic Chemistry is still quite a way from being widespread, but at least they have the raw data for about 100 papers. It hasn't yet conferred any benefit to the readers, but it's a work in progress and a sign that a small proportion of people are prepared to submit complete data.

The issue we're trying to address is that when complete data is turned into a picture, it has a loss rate of about 100-to-1, meaning 99% of it is thrown away in the conversion. You cannot regenerate the data from the picture at all, but you can generate the picture from the data. But the problem is in the uptake, and for people to want to submit complete data, they need to know why it's useful.

MP: Do concerns around data sharing factor in here too?

HR: There are some who have refused to have anything to do with it. They take the attitude of "I'm not giving my data away to anybody." There are also examples of data not being shared for industrial reasons. In chemical synthesis, for example, I knew someone who was a volunteer in a laboratory investigating the mechanism of anaesthesia. They had a leg up on their major competitor because they had a synthetic chemist who could knock up over 100 molecules. For one reason or another, they fell by the wayside, and I asked him if he was ever going to publish the data. He asked his principal investigator for permission to publish them, but he was opposed to their competitors having access to it. No matter how much benefit it could have to everybody else, sharing it was non-negotiable.

The funders have tried to put pressure on. The Engineering and Physical Sciences Research Council (EPSRC) said that with every funded project, they expect the data to be put into a data repository, and for it to have a digital object identifier (DOI). Then the EPSRC was reorganised and most of the people who had issued that instruction were made redundant. The others went off to work for the likes of Elsevier, who started a commercial data repository of their own. That means that the original edict was never followed up on, and word went round that they weren't enforcing it. The Wellcome Foundation is a little better, but with most funders, when push comes to shove, we just want good research. If someone refuses to share their data, that won't stop their research being funded if it's good.

MP: Are you optimistic about the future of data sharing?

HR: In 1965, Cambridge University set up a world database of x-ray crystallography data, which took notoriously long to produce. Around 25 years since the database was developed, journals started mandating entering your data into the Cambridge database. It started with just a few journals and over 20 years, it became ubiquitous, and crystallographers accept that's what they need to do.

There's no other area of chemistry where that's happening. The Journal of Organic Chemistry has run its experiment where one in twelve authors submitted data; the next stage, I believe, is to slowly start mandating it. Another step would be asking referees to comment

on whether or not the data is available in a FAIR form. One of these days, a paper will be rejected purely on the grounds that the data isn't available. I would guess that's around a year or two away. Another journal, Nature, now mandates a data availability statement in each article, where you have to declare how the data is available. It doesn't, however, mandate that the data has to be FAIR.

My hope with the Journal of Organic Chemistry was that if the readership sees the benefit of having data in this form, they will pressurise more journals to do the same. My best prediction is that this FAIR data revolution is perhaps in years two or three of a 25-year cycle. In 25 years' time, we will look back at say, it was obvious that it needed to be done, just like we do for crystallography.

Perhaps in the future, AI quality assurance will be cheap enough that we can still give it a CC-0 licence, and charge nothing at all. You need to think about maintenance, but there are various business models that might be suitable for data in the long-term. What I don't want to happen is that the likes of Elsevier and Springer own the whole shooting match, and charge you as much as the market will bear to sell it back to you. It's up to some people to show that there are other models which won't cost an exorbitant amount. We don't have to be a profit centre for large multinational publishers.