AI 4 Science Discovery Network+

AI4SD Interview with Christopher Gutteridge
03/02/2021
Online Interview

Michelle Pauli
Michelle Pauli Ltd

11/08/2022

Humans-of-AI4SD:Interview-27

AI4SD Interview with Christopher Gutteridge
Humans-of-AI4SD:Interview-27
11/08/2022
DOI: 10.5258/SOTON/AI3SD0241
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	AI4SD Interview with Christopher Gutteridge
Interviewer	MP: Michelle Pauli - MichellePauli Ltd
Interviewee	CG: Christopher Gutteridge - University of Southampton
Interview Location	Online Interview
Dates	03/02/2021

2 Biography

Figure 1: Christopher Gutteridge

Christopher Gutteridge: 'Our goal was to make research available online for free, and we succeeded in doing that'

Christopher Gutteridge is a Systems, Information and Web programmer, part of iSolutions in the School of Electronics and Computer Science at the University of Southampton. He is known for being the lead developer for EPrints and for being an advocate for open data, linked data and the open web.

In this Humans of AI4SD interview he discusses the evolution of ePrints, frustration around citations and the need to standardise data publishing.

3 Interview

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

CG: I did a degree in Computer Science at the University of Southampton, but I got a solid drinker's 2.2, which meant I couldn't do a PhD. The head of department at the time suggested that, instead of doing a PhD, I could apply for a job as a webmaster for Southampton's electronics and computer science department. I applied for it and ended up getting hired. In those days, a webmaster was involved in everything from getting a screwdriver and putting hard drives in web servers to writing articles. I used to have to install and update the machine, build and repair hardware, design templates and programming, as well as most of the artwork, and even writing prose alongside others. But as time went on, these all became specialist jobs.

Eventually, I ended up working on the ePrints project against my will! It was almost completed, so my main job was to make it work with the open archives initiative, metadata harvesting, and protocol version 1.0, which hadn't quite been finalised. I had already written a similar database, but it was designed just for us: it was an early, naive website that had a list of all our research papers. For a long time, I had resisted using ePrints because it didn't do the things I wanted. But later, I got the blessing to start writing version 2 to increase its internationalisation, to make it faster and more configurable. Version 2 was aimed at people like me – systems administrators. Version 3 was aimed at library administrators – I like editing XML and perl files directly, but that's not true of most people who needed to administer repositories. By the time version 3 was developed, many years later, we had hundreds of institutions around the world using ePrints, and it's still active today in many, many places. Our goal was to make research available online for free, and we succeeded in doing that.

Years later, I started a PhD in scholarly communication. Ironically, it never got anywhere because it turns out I hate doing "real" research! What I learned about, however, was the amount of frustration around how citations are handled. People tend to format citations in whatever way they did in their PhD thesis: it's effectively the traditions of your tribe. The point of a reference is to be able to find it again; you need to have enough information to evaluate the quality of the reference. When you read a paper, you should ideally be able to discover all of the papers it cites, and have all of the abstracts embedded in the download to be read offline.

MP: What do you think is holding people back from embracing these changes?

CG: Usually it's a failure of imagination. People don't believe things are possible until they see them. If they can see them working in a local setting, then they can start to see them in a bigger setting, and people begin to ask, "Why isn't this happening?"

Another issue is a lack of motivation. Academics tend to prioritise what gets them funded, and what gets them promoted. Their rewards are not for providing detailed, complex metadata about a paper, they're for doing good papers.

MP: In what senses have there been generational changes?

CG: It's always been the case that a lot of ideas just take a generation to come through. I remember someone from Elsevier telling me that there was an internal war between the old

guard and people supporting open access. Today, people have largely discovered that they can still make money with open access, so they're happy, but there's always going to be this big fear of changing a business model. With larger organisations like Elsevier, it's very hard to turn that rudder around fast.

It only becomes a problem when organisations are protective about things in ways which harm the overall knowledge base. You can do things in the walled garden of companies who possess the data but it's not easy to play with new ways to process and integrate information when it's split between different walled gardens. I would really like to see far more standardised methods for publishing the data associated with research papers. [Note added by Christopher in July 2022: Jisc has just launched a service named Octopus which gives me some real hope in this area!)

On that topic, I have a hypothesis for a visionary system in which, if you're reading a research paper and it has a table of data, that, wherever possible, the underlying data would be attached. While some data is really complicated, much of it isn't, and the majority that's important in research is frequently a set of database tables or spreadsheets. It would be fascinating to be able to embed the source data with enough metadata to use it without having to do any work. If you approach data like that, it also becomes more available to machines to be processed later or for a researcher to instantly overlay it on their own dataset to see if anything interesting pops out.