Interview with Lee Harland

For podcast release
Wednesday, February 7, 2018

KENNEALLY: Computers and human beings generate digital data at the rate of 1,000,000,000,000 bytes every hour. That's more bytes than stars in the average galaxy, and there are more of them every hour of every day. Welcome to Copyright Clearance Center's podcast series. I'm Christopher Kenneally for *Beyond the Book.*

The Big Bang of data in all aspects of our professional and personal lives is sometimes overwhelming. Where it comes to the drug development pipeline, when time matters to patients with cancer and rare diseases, researchers must struggle to keep current with the multiplying volume of information from peer-reviewed scientific research, patent filings, clinical trials data, news, competitive briefs, and much, much more. Machine learning algorithms go a long way to managing the burden, of course. Yet the founder of the leading developer of text analytic solutions is confident that machine learning is not yet able to displace the learning done by human beings.

Lee Harland earned a Ph.D. in genetics from King's College, London, and worked for more than 15 years in life sciences, leading semantic web, data integration, and text mining efforts. He founded SciBite in 2013. While the firm has won awards for its technology, Lee Harland is an enthusiastic advocate of what he calls human curation. He joins me now from his office at the Wellcome Trust Genome Campus in Cambridge, England. Welcome to *Beyond the Book*, Lee Harland.

HARLAND: Hi, Chris, thanks for having me.

KENNEALLY: We're looking forward to this discussion because at this particular moment in the press, there is this notion that the robots are beginning to push human beings out of the picture. But from what I understand from your own work experience, Lee, that that's not quite the case and, in fact, it probably never will be because human curation, as you put it, is always going to be essential, even as we move deeper and deeper into a world of machine learning.

HARLAND: Yeah, I think it can't escape anybody that the machine learning and artificial intelligence is really the top of the hype right now. At SciBite, certainly machine learning is something we use every day. For me, it's a great technology, it's very exciting. But I think it also goes alongside lots of other things that exist, particularly the contribution from scientific knowledge, existing experts in the domain, and how they can bring that to bear on some of these algorithms and analyses that people want to do to gain insight into data.

KENNEALLY: For our listeners, Lee, who don't, as you do, work with machine learning every day, let's define a couple of the terms here. When it comes to data, as I understand it, there are two types, essentially. There's structured data and unstructured data, and it's the unstructured data that's the real critical piece to this, and the most challenging piece.

HARLAND: Yeah, so you can think of structured data as numbers in a spreadsheet. When you know what the experiment was, you know what the columns are, you know what the rows are, you know what the data is. It could be the progression of a certain thing over time, the growth or the reduction or whatever. You've got measurements, essentially. Computers can generally understand that data. But a lot of the information – and you mentioned the growth in scientific articles and patterns in publications and internal documents in companies – a lot of this is just text, and computers don't understand text. They just see strings of letters. They don't really know what it means. So if they see the letters M-O-U-S-E, they don't know it means mouse, and they don't know if that's referring to an animal, to a rodent, and they don't know any other document that mentions other types of rodent. They just don't understand this.

The ethos of SciBite has been to take this, what's called unstructured data, and try and turn into more like this structured information. Trying represent that data in a technical way that computers to say, aha, this document is about a mouse. It's about a rodent or it's about a gene, it's about a specific type of disease. Once the computer understands that, you can then compute, you can then do really exciting stuff with the data that you can't do with just a bunch of raw documents.

KENNEALLY: And you need to do that right at the beginning, right? You need to put it into that kind of structured format so that it's searchable, it's clean, if you will, and you have to do that before you begin the process of using machine learning.

HARLAND: Yeah, I think there's a couple of different approaches. Certainly there are quite a few experiments now where people can just take raw documents and throw them into machine learning algorithms and do interesting stuff. That is certainly not an invalid way forward, there are many experiments you can do. But equally, there are a lot of experiments that you want to do in a data science capacity that require you to really understand a little bit more about the content of the document. So again coming down to machine learning, we often say that – from SciBite, we're actually at SciBite, really big proponents of machine learning. I say we use it every day, and our customers use it.

One of the things we're trying to do is solve the garbage in/garbage out problem. If you're putting lots and lots of random data into machine learning, it's good, but it may not be that good. Whereas if you can go a little bit further and pretreat your data so that it's a bit more structured, a bit more organized, and then feed that to these algorithms, we've seen time

and time again with our customers that these algorithms start performing much better. Again, the quality of what you get out is directly related to the quality of what you put in.

KENNEALLY: Right. We are speaking right now with Lee Harland. He's the founder of SciBite, based in Cambridge, England, and we're looking at how SciBite transforms this data into information and then into knowledge. I guess I want to ask you, what's the secret here? What are the things that are done? I was thinking even with your example of the mouse, it could be a mouse, it could be a rodent, it could also be a computer mouse, and a computer itself, a machine learning algorithm, wouldn't know the difference unless it had some kind of context. Is that what you work on in this human curation? You provide the context, you develop vocabularies and synonyms and definitions.

HARLAND: Yes, that's it. And this all really comes under the banner of an ontology. This is a term that – there's a big irony in the world of ontologies, that the words ontology, taxonomy, vocabulary, synonyms, etc. are not really well defined, even though the purpose of an ontology is actually to define things properly. An ontology, really, is a conceptual knowledge about a particular domain. There is a cell ontology which describes all of the different cell types that exist, and their relationships to each other. For those people who do study ontologies at university, there's a very famous thing they learn in their first year of university, which is the pizza ontology, which talks about how pizzas are split up into bases and toppings, and how those relate to each other. It's really a conceptualization of a particular domain in a computer readable format.

So if we go to the example of a rodent, there are actually a couple of very well-established taxonomies, ontologies of species, which start with things like all animals and to eukaryots and then to birds (inaudible) and then to rodents, etc. What these do is they provide, through computer, a unique code, a unique ID that describes a specific thing, so a rat, a mouse, a human, a dog. But they also organize them, so you know which are rodents, you know which are mammals, you know which are vertebrates. It's all done in a way that, while these things are built by humans, they're completely understandable by machine. That is where the power in using ontologies in this kind of context of unstructured data really is brought to bear because you get to the point where you've taken your unstructured text, you've taken your ontology of let's say, in this case, animals, you've merged the two, and now what you have is a document where, at any point, the computer knows not only that an animal was mentioned, but what type of animal. And of course you can apply this to cells, tissues, diseases, drugs, people, companies, whatever.

KENNEALLY: So that curation team is developing all this. They're disambiguizing – is that the term, Lee Harland? You can tell me. You know better than I.

HARLAND: Disambiguating.

KENNEALLY: Disambiguating, thank you very much. Disambiguating all of these various terms which are important. In the kinds of research that pharmaceutical companies are taking on, knowing precisely which genome, knowing precisely the reaction or the non-reaction to a certain kind of approach or treatment, that's going to matter to drug development. So your curation team does the disambiguation. It expands, as well, the vocabularies. That allows the researchers, then, to work with all of this data in a way that gets them to the results so much more quickly.

HARLAND: Yeah. I think there's two elements to this. The first bit is the who produces the ontologies? I think this comes to the heart of the matter, really, which is that most of the ontologies that we use in this space are produced by a vast community of scientists. Many of the ontologies we make use of at SciBite are produced by the scientific community, funded by both private and public money across the globe. At SciBite we are consumers and contributors to those. You have initiatives such as the Allotrope Foundation, which is, again, (inaudible) pharmaceutical companies and other partners who've come together to develop these standards for pharmaceutical sciences and processing of drugs. I think the power is in the openness, the fact that they are done in the public domain, they are free to use by everybody. Because this gets to the heart of why you want to do this. It promotes data interoperability, and actually the ability to do these experiments.

In my previous life working for a major pharma company, we had issues wherein two different suppliers would provide us with data that was described in – let's say they both had the concept of the mouse or human data in there, but it was described using two different proprietary ontologies, and it's very difficult to work with because it doesn't connect you to any of the public data or any other data.

The first half of the answer to your question is really that the ontologies we work with aren't the result of one, two, three, four people. They're the results of thousands of experts, everyone contributing a tiny little bit of knowledge to an overall coherent map of a particular set of cells, tissues, diseases, whatever, and then the power to be able to leverage that expertise in a computer-readable format is incredible.

KENNEALLY: Indeed. I was thinking that the notion of collaboration and community, that's essential to scientific research because the work you do, Lee Harland, in your lab in Cambridge, and the work that I might do at my lab in Boston, Massachusetts, we want to share that information, we want to build on that. And then someone else will come along to build further on our work.

HARLAND: That's exactly it. So proprietary ontologies, while they do have a place, really the power of this big data world is being able to – when you describe a mouse and I describe a mouse, if we're describing that to computer, we want to use the same description so the computer can see that we're talking about the same thing.

The other half of this, though, you mentioned the disambiguation side, so when people like many of the folks at SciBite and many of the people on the team at Genome Campus here in Cambridge in the UK, and of course around the world, are building these ontologies. The primary aim is to represent the things. So if you're working with ontology of animals, your primary aim is to represent the list of all the animals and organize them into their correct taxonomical categories. They're less focused on the different applications that people might put those ontologies to, because of course you don't know, people can do lots of different things with these.

One of the things that we want to do, and what we do at SciBite, is take these ontologies and use them to find those concepts within text. Actually you raised this idea of disambiguation, that's actually quite hard. The example I always talk about is hedgehog. We all know the hedgehog is this nice little spiky animal that ferrets around in the garden late at night. But to many scientists in life sciences, hedgehog is actually better known has the name of a protein which is absolutely critical in cell division, which, of course, is the major process involved in cancer. So when you say hedgehog to a life scientist, particularly in molecular biology or human genetics, they're much more likely to be thinking about the hedgehog gene or protein, and not the loveable animal.

When you are trying to apply ontologies to text, you've got to do it in an intelligent way. When you see the word hedgehog, you've got to build systems that say right, OK, this could mean one of two things. So I'm going to have to do a little bit more work to work out what it exactly is. I'm not going to annotate it as the hedgehog protein unless I really think it is, and similarly I'm not going to annotate as the hedgehog animal unless there's something that tells me that it is the animal. That's disambiguation. When you apply ontologies to text, you can't just do it in a naïve way, you've got to have a layer of intelligence that really tries to understand exactly what that word means.

KENNEALLY: Because when you apply the machine learning algorithms and so forth, what the researcher is after is scientifically interesting results, and all that work you've just described is going to lead us there.

HARLAND: That's exactly right. Of course, this isn't a battle of AI – machine learning vs. non-machine learning because actually at SciBite, we use a lot of machine learning techniques to build the disambiguation inside ontologies. So these things are not polar opposites. This is an ecosystem of technology where applying the right technology at the right time is what you want to do. So you can use machine learning to help build these really deep layers of disambiguation and rules for applying ontologies. Then, of course, you can apply the ontologies to extract out really cool data, and then you can push that back to machine learning to do some nice analytics. So I think the point is that these things are not opposed to each other. It's more the two combined give you an incredible power that up until a few years ago was just not available.

KENNEALLY: And rather than being afraid of artificial intelligence, the work that you're taking on at SciBite, Lee Harland, and the work that goes on around you at that Genome Campus, it must be thrilling to live at this particular moment in science when the kinds of results you can get combining the human curation and the machine learning are just going to lead us to cure diseases and to solve problems that were beyond our ken in the past.

HARLAND: I think it certainly is an exciting time. In my team I've got a couple of people who did Ph.D.s in machine learning and artificial intelligence 20 years ago. They're bittersweet because they're loving the fact that their skills are now seen as very important. They're also annoyed that for 20 years, people just ignored them and didn't think they did anything that interesting with their lives. But now of course they're people of the year. But this is where the hype comes in, so I think what you see a lot now – and this is expected of any new technology, we saw it with Semantic Web, we've seen it with many other things – that the hype has almost gone ahead of what's real and sensible.

Our experience has been that if you cut away the hype and you look at what we're doing, what our partners in different companies are doing, what our customers are doing, the actual use of machine learning and ontology-based technologies, you can now do some quite powerful stuff. It's not yet giving you the magic phone app where you type in your symptoms and it suddenly gives you the cure to your disease. But little things like – I guess I can talk with you about some of the examples that I can tell you. Little things, the stepping stones to the broader picture. So I think my message is that right now machine learning isn't this magic bullet that in two years' time is going to solve all our problems, but it is a technology to be taken seriously, and actually has had some real impacts in some of the things we do.

KENNEALLY: We appreciate your perspective, Lee Harland, and certainly have learned a great deal y chatting with you today. We've been speaking with Lee Harland, he's the founder of SciBite based at the Wellcome Trust Genome Campus in Cambridge England. Lee Harland, thank you for joining me on *Beyond the Book*.

HARLAND: Thank you very much.

KENNEALLY: *Beyond the Book* is produced by Copyright Clearance Center, a global leader in content management, discovery, and document delivery solutions. Through its relationships with those who use and create content, CCC, and its subsidiaries, RightsDirect and Ixxus, drive market-based solutions that accelerate knowledge, power publishing, and advance copyright.

Beyond the Book co-producer and recording engineer is Jeremy Brieske of Burst Marketing. I'm Christopher Kenneally. Join us again soon on *Beyond the Book*.

END OF FILE


BEYOND
the BOOK
A Copyright Clearance Center Podcast