Are Publishers Ready for the New Readers?
Recorded at STM Week 2019, London

For podcast release
Monday, February 24, 2020

KENNEALLY: While the public may marvel at machine-generated output from Siri and Alexa to their questions about the world, publishers now understand that producing the input to help such machines form their answers is an attractive, forward-thinking opportunity.

Computers, however, do not read in the same way as do humans. Savvy publishers recognize the types of adjustments that will cater to this new machine reader and then make systematic changes across their repertoire, or perhaps in a specific subject area, to maximize results.

In today's discussion of publishing for machines, I'm very happy to welcome a panel of distinguished technologists who will take us through steps to prepare content for computer consumption. We'll start with our panel introductions. From the very far end, I want to welcome first Lucie Kaffee. Lucie, welcome. Lucie is a PhD researcher at the University of Southampton researching in the Web and Internet Science Research Group. She is currently a research intern at Bloomberg in London and a Newspeak Fellow. Her research focuses on multilinguality in structured data, particularly in supporting low-resourced languages in Wikipedia. Previously, Lucie Kaffee worked at Wikimedia Deutschland in the Wikidata team and is currently still involved in Wikimedia projects.

To her left is my colleague from Copyright Clearance Center, Tom Morris. Tom, welcome. Tom has spent more than 20 years in IT as a principal architect in publishing systems integrations projects, before becoming CTO at Ixxus and subsequently senior director, engineering, at Copyright Clearance Center, which acquired the London-based startup in 2016. In this role, Tom has led infrastructure and development teams, but has given special attention to content and knowledge organization systems.

To his left is Sadia Shahid. Sadia, welcome. Sadia is director of strategy, growth, and partnerships at Wizdom.ai, an AI-powered startup from the Oxford University Software Incubator that was acquired in 2017 by Informa. Sadia is part of the founding team at Wizdom.ai, which interconnects billions of data points about the global research ecosystem using artificial intelligence, machine learning, and natural language processing.

And our final panelist is directly to my right here, Andy Halliday. Andy, welcome. Andy is senior product manager at Springer Nature here in London. His work is focused on the development of content services and the future of content. This work includes looking at how new technologies can solve user problems by improving efficacy and usability of scientific content.

So obviously a very well qualified group to discuss this issue, but I do want to start, Lucie Kaffee, with yourself and your very important work with Wikidata. We are in a group of publishers here, but it's interesting to think about Wikipedia as a publisher in a very different way, but one that perhaps has learned some things that you can pass along to this group. It's important to recall that Wikipedia has its origins back perhaps not even in Web 1.0, but maybe Web 0.75 or something like that – very early days, at a moment when the reader was the human reader. That's really what counted. Today, obviously the human reader still counts for a great deal, but the machines are reading. Describe for us the challenge this has meant for Wikipedia and how Wikidata has begun to address that.

KAFFEE: As you know, Wikipedia is mostly for human readers, so none of the data was reusable outside Wikipedia. And I think this was one of the main incentives to say – so there are two main incentives. One was the problem that machines couldn't reuse our data, so we would be very limited in what applications we could support with Wikipedia. On the other hand, also inside Wikpedia itself, there are multiple language versions. So to cross the language barriers between them and support also lower-resourced languages, we said there needs to be one central database to store all this information.

That's when Wikidata started. So it started to be a support tool between different language versions of Wikipedia and now became a widely reused resource for a wide range of applications. For example, Google will use it to display data. It's searchable. It's usable. We can create it in a wider sense than just Wikipedia.

KENNEALLY: And the machine's role in all of this has grown over time. You were telling me an interesting statistic as far as machines who are creating content and using content. It's far greater than any human being's.

KAFFEE: Exactly, so that's a very important thing. Wikidata, same as Wikipedia, is completely community-contributed, so there is no – I don't know – paid person to integrate data. But what happened over time was that Wikipedia edits, so people writing on Wikipedia, is relatively tedious. It's not automatable. So what happened after a while was that Wikidata now has per day and per minute more edits than all language versions of Wikipedia together, because we have auto-

created content. So there would be automated tools that would go through the internet, extract structured information, and import it into Wikidata, which made it a vastly growing application.

Then again, what happened after a while was in the beginning, people would take information from Wikipedia and import it into Wikidata. So slowly, slowly, what we are doing now is that we take information from Wikidata and display it on Wikipedia, because there's just so much more data in Wikidata just because we can automate the process of how we create the content.

KENNEALLY: Has this changed at all the relationship that Wikipedia has to its editors, to those who are contributing the content?

KAFFEE: It's very interesting, because quite rightfully, the editors of the large Wikipedia, so English or German, are not very happy with the idea to integrate the data into Wikipedia, because it seems to them a bit magic, right? It's really hard to make people understand what's happening and how it's happening. And I think that's a very important thing to do whenever you work with structured data and AI and all those things, that you make people understand and you make your community and your users understand what's happening. While the small communities are really excited about the fact that now they can use data that is constantly maintained, because we have one central place for everyone to maintain the data independent of language.

KENNEALLY: And I think we'll hear later on that the effort to keep current with all of this is going to change the relationship that publishers have with authors, with their contributors. So you started to touch on that, but I wonder whether there's an effort underway to provide direction for contributors so that this problem isn't always a looking backward issue, but you really have kind of caught up and can move forward.

KAFFEE: Yeah, so I think that's a very important idea of being flexible to integrate those methods in what you're doing.

KENNEALLY: OK. Well, Andy Halliday at Springer Nature, we heard earlier today about that exciting project from the spring when you published a book that was written by a machine on lithium batteries. It got a lot of attention at the time. I had a discussion at Frankfurt, actually, for STM with one of the leaders of that project, Niels Peter Thomas, and we talked about some of the issues that were raised around ethical questions and the roles of authors and publishers and peer reviewers and so forth. It's quite fascinating. But give us a sense of what level of excitement there is around all of this within Springer Nature. I'm fond of one to 10 scales, right? So

if one is a baby taking a nap and 10 is a kid on Christmas morning, where is Springer Nature on that? What number would you say?

HALLIDAY: That's a good question. I think it's probably somewhere in the region of six and a half to seven. I think from the perspective of what we tried to do and set out to do with that particular example was to prove that we could do it. The content that is generated or that the book is generated from – all peer-reviewed content. And actually, what we found was it was quite an interesting process to work with subject matter experts to make sure that what was coming out of the system and what was coming out of the algorithms and what was being written was actually true to what was being fed in. I think that was very much an academic exercise. That one was to prove that we could do it.

There are plans afoot to continue to try and look at this. It's not something that we're going to be rolling out 10,000, 15,000 books in one year, but there is active work going on to try and continually improve what that looked like, but also what we can do in the future.

KENNEALLY: And it's not only about writing, authoring, but it's about reading. That's what we're talking about right now. So that provided you with some insight to Springer Nature about how readable your content was, I would imagine.

HALLIDAY: Yes. I think one of the big things that we found, and it's generally across our corpus – a very basic thing about machine readability is having the availability of structured data for all of your content. Now, we found through that process that it's not always there, so we've embarked upon a much bigger program to try and start provisioning full-text structured data for all of our content so we can then actually widen the information that goes into any kind of summarization book in that case or summarizing journal content for a different purpose. But it's actually allowed us to – realizing that having full-text XML basic, having good ontologies to describe the content well, having it well tagged and consistently tagged so that actually machines who read that content can make linkages between things which could be so far apart as 10, 15 years and across various different geographical boundaries.

KENNEALLY: Right. We were speaking with Lucie about the past, about the origins of Wikipedia more than 20 years ago. Are we still talking about something we were also talking about 20 years ago, which is XML – it's really still something that people are working on?

HALLIDAY: Unfortunately, yes. XML still forms the basis of all of the content that we publish. Obviously, we comply with (inaudible) standards for content. We have

our own proprietary XML standard that we use internally. But it's all very much based upon XML.

Having said that, that is a first step. It's kind of making sure that we've got that. We have examples of content from a long while ago which we only had in PDF. So trying to have a process of pulling that out, extracting that, making it structured, and actually making it available to machines to read it – that is something that is a process that we're working through.

KENNEALLY: Sadia Shahid from Wizdom.ai, you were telling me that this is a world we exist in where data is far from perfect, and Andy is alluding to that. Indeed, we've heard that already from Lucie as well. Talk about that as a challenge, that as far as we have come with the web, with technology, there's a journey yet ahead with making our data as good as it ought to be for the future.

SHAHID: So traditionally, publishers – historically, they were into publishing papers and journals in hard-copy format. And then as time passed, we've gotten used to it being put up as a PDF, but the format has pretty much remained the same. And more recently, with the advent of technology and making it more machine-accessible, we've moved into formats such as the HTML and the XML, but still it's mainly the PDF format that has been transformed into the XML. What is still lacking is good-quality metadata, so information that describes the content, good-quality identifiers that describe that – the entities, the authors, their affiliations behind that data, as well as the content in itself. So if we're talking about an age where Alexa and Siri are answering questions and where pharma companies are mining through millions of papers to develop new drugs and treatments for diseases like cancer and Alzheimer's, we need to have it in a format that it makes more sense to the machine.

Just talking about why NLP is so important and why machine-readable content is so important, I'm going to give you an example. Let's say if I were to make a statement like Chelsea is on fire today. Am I talking about the football club and that they're performing very well today, or am I talking about the place Chelsea, and is it literally on fire? Or am I talking about a person Chelsea, and she's giving a great performance, or is she on fire? So that kind of context – unless I go beyond syntax, I go beyond and discuss the context in which I'm talking about this statement, a machine would not be able to interpret it in the way I as a human can.

To make that more comprehensible to machines and to help the bots do systematic reviews and literature reviews in an automated fashion, we need to make XML in a format that is more easy to understand by tagging – for example, one way of it could be to tag what are the exact definitive claims coming out of this research?

What are speculative claims? What does it affirm? What is the hypothesis? What research is it refuting – in a form and in language that is comprehensible to that machine. That is where I see the future as. That's the next step – taking it beyond the XML and making it more available and accessible to the machine.

KENNEALLY: Right. Tom Morris, we were talking about XML, and this is a great point. The question I put to you was are we still talking about XML, and you said, yes, we have to. But the reasons have changed. Catch us up on that. Why are things different? When we first heard about XML, it was a format for publishers, perhaps even created by publishers, to help them with a certain challenge. Today, the challenge has changed.

MORRIS: Yeah, so XML has been around for a long time, and I think the primary motivations for the machine readability were quite straightforward transformational needs for format-shifting to move toward single-source publishing, for indexing purposes for improved searchability, for kind of very pragmatic reasons. But the same reason it was valuable to have readability in the past is the same value we have now, but for a different purpose. The purpose is machine readability.

I think there's a statistic somewhere that talks about machine learning endeavors – 80% of the cost of those initiatives are clean data and making sure that the data and the structure and XML are there in the first place. So I think one of my interests is looking towards ways of incentivizing people to get that 80% with reduced friction, and maybe the sensible compromise is with the XML.

I think we were talking earlier on about how Google rich snippets was a great example of how mainstream authors who didn't require or didn't value the fact that it was machine readable, because they're just targeting human readers, made something machine readable through the addition of invisible tags, the microdata, that means that Google suddenly makes itself look very smart by smartly presenting a piece of data – oh, it's about a hotel with a rating and so forth – improving the clickthrough rate and improving commercial viability of the original author. So that was an interesting incentive. And sometimes I think the incentive could be much more pragmatic.

For example, we partner with Fonto. It's an XML editor. That reduces change management. The author doesn't even know they're authoring XML. So I think that's kind of where a lot of our effort is focusing in, which is improving readability for machines through very pragmatic user experience-focused reasons.

KENNEALLY: Right. And we heard that this morning with Bill Kasdorf's presentation about the pivot to the practical, and you're really addressing some of these practical

points. It's interesting. We're supposed to emphasize tools here, but human beings keep coming up. It's a surprise to me as a non-technologist, because I think, well, there's going to be a technology solution, but really it's a human solution. Everyone here is nodding for that.

Lucie, add to that conversation about – again, I want to push you on the role that the authors, the editors can play in making this material more accessible to the machines.

KAFFEE: Yeah, I think that's a core part of Wikidata. As I said, it's all community-contributed, so none of the data is magically just appearing. There are now advances in research where, for example, there's something called link prediction, where a machine learns from an existing knowledge graph and tries to say, oh, that is a statement that's possibly lacking.

But in the end, our AI tools are not at a stage where it understands the world as a whole. So in the end, we can just reproduce what's already there with AI, so we're just learning patterns. So in the end, we always need humans to contribute to the data. And especially when we want to use machine learning, we need a very clean and well distributed data set to actually make them understand how the world works and how things happen – what's a bias in the data, which is a really big topic, I think, in this regard as well.

So in Wikidata, as I said, we have all those bots that import data. But all of those bots are written by humans, and they work over data that is also written by humans. For example, they will import facts from Wikipedia, which in the end is also human-contributed. So I think we're not at a stage where we can just leave the machines alone to just play.

And on the other hand, as well, everything we do then is also for human readers, right? When we were talking about this earlier, all the structured data we collect and use for some models and the output in the end is also for humans. When we summarize for research purposes, discoverability of new papers, of new topics, all of those things are not just for machines to train themselves, but they're for a researcher in the end to go through them and integrate it in what they're working on. So I think it's always very important to have an eye on that when we talk about structured data as well.

KENNEALLY: Right. As we take this material from the human reader to the machine reader and back again, it's kind of a virtuous cycle –

KAFFEE: Exactly.

KENNEALLY: – and one that, as you point out, humans are playing a greater role in. Yet it is to teach the computer again. So it's never really toward an end. It's a continuous loop, if you will.

Sadia Shahid, with your experience, talk about that role that the human beings play in teaching the machines to do this job better. I believe that you feel it's important to keep up with ontologies, and that to address this particular audience of publishers, keeping up is a real challenge with ontologies.

SHAHID: Yes.

KENNEALLY: And you should probably tell everybody what we mean by that. It's kind of a dictionary.

SHAHID: Yes, so that's basically a related set of terms and the scientific denotation of that. How do you define those terms and how they're interrelated? Basically like a lexicon that may be specialized.

Interestingly, when Lucie was talking about bias – and I know you do a lot of work with different languages – when I was thinking about this talk and this topic, one of the less-talked-about discussions are around ontologies and content beyond the English language. Why I want to emphasize that is because if you look at the recent trend and where in the world research is happening, China has surpassed the United States in global research output. If you just talk about one topic, which is, for example, if you take artificial intelligence, not only has China exceeded the quantity of papers that they're publishing, the research that is happening over there in artificial intelligence, but also they've surpassed in terms of quality of their content. The top 10 cited publications in artificial intelligence this year have come out from China.

If you just take that particular subset – research in China – I looked up one of the biggest collectors and aggregators of the data, (inaudible) Corporation. If you look at just Chinese publications that they have in their database, they've got around 43 million articles in Chinese journals that are about 8,000 journals that they have. Only 5% of that content is English-language. That bit, that huge significant portion of research that is happening right now, we do not have the language to describe that. We do not have that content readily accessible to be indexed in our databases. If you're talking about Web of Science, if you're talking about Scopus or any other systems out there, the publishers are not able to look up as to what research is happening over there in those papers – what are the emerging areas being talked

about? What is the latest findings coming out of that? Just because we don't have the vocabulary to define that.

So beyond English language, we need to look into how natural language processing can play in different languages, particularly in Chinese, and develop ontologies that are transferable, and machine learning and artificial intelligence can actually play a part in that. So automating the tagging of content and defining what are the terms and what are the topics and keywords being discussed in a particular paper – if we're doing that smartly for the English language in the first place, we could also kind of transfer that knowledge across and translate that across different languages.

KENNEALLY: Lucie Kaffee, I have to ask you to join that discussion, because that's your area of specialty is multilinguality in data, and particularly for what Wikipedia refers to as low-resourced materials, which are just languages beyond several Western languages. You've had some interesting experiences around that, and again, it points to the important role of community in all this.

KAFFEE: Exactly. So that's why my research also focuses on Wikidata, because the knowledge graph in itself is language-independent. Let's say we have an entity – we have something that describes London. That will be not London as a tag, but it will say Q64. So it will give us a language-independent version, where then we can translate it in all kinds of languages. So we have readily and on the fly always the option to switch between languages seamlessly and easily, because the statement London is the capital of the UK is completely translated. Like each part of it is language-independent and then can be referred to in different languages, which on both sides works really well, because on the one hand, of course English-speaking researchers can access Chinese information, but on the other hand as well, Chinese researchers can access English information.

So there is this giving and taking between different language communities, which we've seen especially in Wikipedia a lot, where as I said, the low-resourced communities are very happily taking in the information that is maintained and created by other language communities. And it works the other way around, so we get access to information that originally is only accessible in, let's say, Swahili, but then suddenly we can have that in a format we can access as well, which forces multilingual and across-nations collaboration as well, which is something that especially in research is really important as the world grows and becomes more complex. I can see this especially in computer science, where you have experts in the same field across different countries and different languages, where a seamless integration of the research content is really important.

KENNEALLY: Andy Halliday at Springer Nature, your experience with developing these ontologies, keeping up with all the changing sciences and so forth – what's the experience like? Human beings play a role, but I think in your example, it's also a machine learning challenge.

HALLIDAY: Yeah, so I think there's a number of challenges we've faced at Springer Nature. From the perspective of the publisher becoming what it is now by multiple mergers over time, there's been various different ontologies that we've had to try and agree on. What are we saying is the canonical way we want to describe this? Now, that's been a challenge.

So one of the things we're trying to do with that is actually develop a new Springer Nature taxonomy which is structured fairly differently to what we've had in the past. There's elements which are discipline-level which are not going to change – so you've got chemistry, you've got physics. But then the kind of sublevels of that are much more granular, and this is where we get down to individual pieces of content being tagged with a very specific scientific term or an entity, or maybe even a chemical structure.

But the problem there is everything is moving so fast – we've also got publishers in China, but other English-language publishers around the world are publishing new things, suggesting that there are new types and new subformats or subdisciplines of an overall broad discipline. How do we make sure that we're reflecting that language effectively?

So part of the work that we're doing – we have a knowledge graph ourselves which we're pulling pan-publisher content into and using machine learning across that to pick up when new ways are being used to describe content that is similar to the content that we have, and we have the multiple linkages with the knowledge graph to do that. And then we are at the minute still using humans to ratify that that is actually the same thing or it's something which is – that somebody's using a different way to describe something that actually is never going to take off. It's something that is here. But ultimately, we're starting to train a machine to start looking for new concepts and then allow us to then bring them in to keep our ontologies up to date and keep our content as current as possible.

KENNEALLY: Right. Tom Morris, not every publisher is perhaps – and maybe you'll argue with me, Andy – as well resourced as Springer Nature may be. So there are some tools available that can get people pretty far along the road. I believe there was one from Amazon you were telling me about.

MORRIS: Oh, yes. So the concept of entity extraction I think is important to think of here, because it's become much more of a commodity service offered by lots of people. They could be used to generalize out of the box – within seconds, you can be extracting valuable information from a document on a generalized data set. So it might not reflect your specific domain, but you can train it. But that low barrier to entry is really pivotal here.

I don't have the same breadth of expertise, but I recognize some of the problems just talking about ontologies, because we typically have a top-down approach to defining those. We have your expert-identified large categories and then slowly break it down, but then there's rigidity in that, and there's also scope for bias. So I think what I'm hearing here is that there's also tools available to use machine learning to surface it from the bottom up – identifying the granularity and maybe weighting them based on frequency to kind of get a better value proposition.

But don't forget that humans can help us here. This is kind of a two-way thing, and there are many, for example, scientific vocabularies out there from the likes of SciBite and so forth that have huge amounts of valuable, very frequently updated, human-curated data sets that can then go in and reinforce the machine's understanding of a document. So the two go hand in hand, and I think they always will do.

KENNEALLY: Yeah. Before we get to some questions, I want to go back to the beginning, which is the spark for this conversation was recognizing that voice is now playing an increasing role in the consumer experience with technology, and we can expect to see it begin to creep into the business world, the research world, the university world. I wonder if you could all just address how this kind of work that is underway is going to prepare us for that world, where voice plays an increasing role. I believe we are still at a point where publishers are still mostly with keyboards in front of them, so we're a bit ahead of the curve. But, Andy Halliday, is voice part of the considerations that you have at Springer Nature?

HALLIDAY: It is. There's some work that we're doing which is around auto-summarization of content and snippets which can be personalized to an individual. One of the ways that we recognize people want to access that is by asking a voice assistant how to do that. Now, being able to give something back which is not going to be an entire scientific paper or an entire book chapter, which is going to be way too much, potentially, for what that person wants is tagging content down to a level which is a paragraph or a particular sentence or something which is actually going to be able to be pulled out to then summarize that content back to the user.

Now, that is one element of it. I think there's another element that publishers, I guess, need to think about as well. If you say it's fine to have the voice input – take an example of a chemist working a lab, hands covered in nasty chemicals with gloves on, obviously. You're asking Siri, Alexa, Google Assistant, whatever it is, to give you a snapshot of something or give me this equation or show it. At the minute, the follow-on interaction from that is going to be probably still having to type or touch a screen. So how do the publishers then think, OK, we've got the input to request the information, but how do we get that output in a way that can be used without somebody having to use their hands? So I think that's another challenge.

But getting the content in the first place – it's something we're thinking about, and getting that real granularity of tagging down to really kind of content-based levels is one of the ways that we can then look to start generating this extra content.

KENNEALLY: Tom Morris, that role that voice plays – it's got a real attraction for the consumer of the content. It's going to be a real challenge for the publisher to respond to all the expectations that the consumer has. But you have a vision for how it could play an important role in educational publishing.

MORRIS: Yes. So what we're seeing emerging is that it doesn't have to be an either/or, like do you choose to use the web textual interface, or would you rather use a voice assistant and use that alone, but seeing a fusion and how in real life, you can use that together with the text as kind of a coaching aid. Which takes us back to our initial point of how do you publish for the machine? If you had an educational document that's taking you through a series of quizzes and statements and things to learn, how do you provide context in that text? How do you give cues and clues to the machine?

At this point, it'd be interesting for the machine to ask you a question whilst you're learning as kind of a coach and say, I noticed you didn't reduce that equation any further. Was there any particular reason for that? Do you think it can go further? And you can carry on and go, yes, I think I can, and it might take you to a reminder where it can remind the student how to reduce the equation even further. So much more of a hybrid between text. But that does require the initial author to give valuable cues to the machine of when you can start having a discussion.

KENNEALLY: I want to thank our group today – Lucie Kaffee with Wikidata and a PhD researcher at the University of Southampton, my colleague from Copyright Clearance Center, Tom Morris, Sadia Shahid, director of strategy growth and partnerships at Wizdom.ai, and Andy Halliday, senior product manager at Springer Nature. Thank you all very much indeed. Appreciate your time.

(applause)

END OF FILE