Andrew Lih in conversation with Deb Howes Apr 26, 2023

Deb Howes
What could happen in the future? So much hand-wringing is happening right now, but I think it limits our thinking about, hey, wait, you know, like, AI could be a really great thing. I mean, just to take the obvious example that you're already, I think, thinking, which is like, if every museum put their most essential

and even non-essential things into Wikim data, Wiki, whatever. What would that mean? What could we do as a result of that? I think nobody's really talking about that. And if that doesn't appeal to you, we can go in a totally different direction, but I would love to hear Andrew Lih talk about that.

AI Hype is real

Andrew Lih (01:35.995)
No, that's a good question. And something that most people don't think about is, I absolutely do think that the hype around AI for the last, let's say, six months is real. I mean, what does it even mean to say the hype is real? No, I think there's a there there. Not since the 1990s when we saw the web become such an influential phenomenon.

If you remember some of the sayings back then, 1990s, like, today, every company needs to be an internet company. Or today, every company needs to be a web company. And at the time, people are probably like, what do you mean? I'm a garbage disposal company. Like, why do I need to be on the internet? But we know today, of course, everyone's on the internet, right, whether it's social media, your website, how you transact business, how you book your garbage collection, how you inform them your way for vacation, all these things, it's all through the internet.

Andrew Lih (02:33.258)
The hyper on AI is absolutely legitimate, that every companies can be touched by AI in a significant way. And if it hasn't already, so I was really enthused to hear about ChatGPT. It is almost a household word at this point, even though there's many aspects of AI that are not ChatGPT, but it's kind of, that's the carrier wave, right? The website was the reason why people said, oh my God, I gotta pay attention to the internet or email, whatever you wanna say. ChatGPT is the reason why we should pay attention to AI because...

AI is one step away from magical

It is one tiny step away from magical, the same way that the internet was one tiny bit away from magical, and that I can look up anything at any time. And with AI, it's like, these things can solve problems I never knew computers could solve. So I do think that the AI phenomenon is real. Now, why does that matter in terms of museums, education, and Wikipedia? The reason why, because all these AI systems depend on really high quality training data. And if you look at the data, it's

Wikipedia is at the front line of AI training

Almost every AI system in the world, Wikipedia is right at the front line in terms of the corpus that they're trained on. Think about it, where would you go to get the most accurate and expansive information and knowledge on the internet today? Would it be from a site that you have to pay for it? Well, you have to pay for it, right? Or would it be a site that has lots of advertising? Because you'd be worried about the advertising influencing what you find there. Is it a liberal new site? Is it a conservative new site?

WP is mulitlingual

And if you think about it, and the more we look around, Wikipedia is in fact the best resource on the internet in multiple languages that has been established over the last 20 years as the go-to place for answers, whether it's, hey Google, what's the capital of Montana? Or it's, hey Alexa, who won the World Series in 1973? We go to Wikipedia all the time, whether you know it or not, right? And that's why it's so important for the AI revolution that Wikipedia is right there as the main corpus.

that these AI systems are trained on and depend on. So if you look at these, what they call the common crawl, the library that allow these AI systems to start with, Wikipedia is right up there with all these other sources that you find out on the internet, whether they're public domain books or library collections. Wikipedia is really the best place to start for most of these things.

Deb Howes Studio Com (04:52.166)
So you may not know the answer to this. And by the way, I'm totally on board with your opening, Salvo, bravo. You're right about the sources online that have, let's say gates, like you have to pay money. Well, what is it about like NYPL website or even Google Books or like why aren't those sites

Deb Howes Studio Com (05:22.154)
important in your mind, or maybe they are and we just didn't get to it, but like, how does the training systems favor one thing over another?

Libraires as AI training grounds

Andrew Lih (05:31.530)
Yeah, that's a good question. I mean, you're probably thinking like, well, libraries do that function, right? Yes, though libraries, in terms of what they can provide, are in fact still working within the confines of things like copyright. So if works are expired or they are public domain, then that's great. You can go to your public library, Library of Congress, National Archives, download that, and an AI system can ingest it. But.

If you're still working within the confines of authors that are still alive, or it hasn't been 70 years past their death date, those are copyright encumbered. So you can't actually scrape that content. You can't get free access to those things. So that is a big challenge. What you find on

Wikipedia is because it's all free content that's been contributed by folks and freely licensed, you get the expansive.

human knowledge on Wikipedia in a pretty good way. But for other, if you're just doing it the raw primary sources and primary texts, you are somewhat limited by whether it's still copyrighted or it's available freely.

How digestible by AI is Museum info?

Deb Howes Studio Com (06:43.162)
Okay, so let's just assume that the primary sources you mentioned are past copyright restrictions. Is it an efficiency issue with like, because for example, the Met has hundreds, maybe even thousands of their past catalogs open and available, and some of them are even in a format that a lot of other museums use.

developed by the Getty, which is intended to be open access. Are those formats easily digested and understood by the machine learning systems? Or is there something missing? And I just, you can compare it to Wikipedia that makes it harder or less easy to learn from or anything like that.

Western Bias

Andrew Lih (07:34.122)
Yeah, I think the best practices of respected institutions like the Met, Smithsonian, MoMA, Louvre, British Museum are great. Though you also have to recognize limitations. Do they really fully represent the breadth of human civilization and arts and experiences? And actually, they don't. They're still very much Western in their scope, even as they try to embrace things like South Asian art or East Asian art.

It's still very limited, right? So it's still a very specific lens that you see that content through. Also, they are somewhat limited in the interpretation of those things, right? So the major topics will get a pretty deep treatment. But on the long tail of different topics, they might not be getting great treatment. And you'll often see on Wikipedia, because it's such an international community and having thousands and thousands of volunteer contributors, that you can actually get

better coverage of things related to East Asian art or Aboriginal art than even well-resourced institutions because you actually have a greater scope of contributors than those institutions.

Deb Howes Studio Com (08:48.838)
That's a really good point. I hadn't thought about that so much as like the, at a publication level. I know you've done a lot of experiments at the Met and probably Smithsonian too, about the bias inherent in the collection data. Do you feel that, do you feel that if, you know, let's just say, for argument's sake, if every museum could focus on evening out.

the data in their collection so that Western and non-Western data was equivalent in terms of its detail and authority and everything else in Wikipedia. Do you feel like that would be enough? That's the best and most important thing for the museums to do, as opposed to, let's just say, publishing more catalogs that would even it up.

and having those accessible online, maybe not copyright free from some point of view.

Can the bias be corrected?

Andrew Lih (09:52.106)
Yeah, there's still a lot of work to do in many dimensions. So it's not even just the supply of primary source material. But even the metadata standards we have are still very much biased towards Western art. So if you look at the well-known thesauri vocabularies or the metadata standards, whether you're talking about Getty, AAT, or ULAN to describe artists' names and artists'

or something as detailed as like what the ICOM has in Europe for religious art. They're still very much focused on Western art. So something we've seen quite a bit is when you talk about Eastern and South Asian art, there are not great metadata standards for describing those types of things that we need a deeper, deeper scholarship to even find what is.

Importance of metadata

done well and what is missing. So there's a lot more foundational work that needs to be done just around the meta data that we use to describe that.

Deb Howes Studio Com (10:59.823)
Do you have any recommendations about how museums might go about doing the work that you described, which I agree is really essential?

Andrew Lih (11:08.522)
Yeah, it's a good question. I mean, you'd like to see entities or international entities like UNESCO and these types of folks support more work like this. And to be fair, you have had multiple folks in the, I would say, what we sometimes call open glam. So the open content in the heritage and cultural sector recognize these are shortcomings and try to build up these areas, but it's tough, right? Because...

Difficulty of solving the metadata problem

You need a lot of resources, you need more scholarship, you need a more active community to coordinate around this. And unfortunately right now, you don't have that much being put into this effort to come up with good metadata standards outside of, you know, the well accepted areas for museums that have resources, right. And at least right now, the number of museums or NGOs in

let's say Asia, for example, to spearhead this kind of work, there aren't a lot of them doing that, right? So that's something that you'd like to see is more coordination among entities to better develop these metadata standards.

Deb Howes Studio Com (12:19.198)
Um, you know, uh, I could just digress for a second, which is in that before internet time in the eighties, I was part of a group of other museums, uh, representing the art Institute, um, in Chicago that, uh, saw an opportunity to have one, uh, art collection cataloging system that we all subscribe to. And in conjunction with that.

Deb Howes Studio Com (12:49.778)
we would have, we made a consortium called Amico, which was the art museum interchange cooperative, I think, which allowed us to all share images with each other as a kind of educational effort. This is way before what is now known as Art Store came into play. And I,

Andrew Lih (12:52.247)
Mm.

Andrew Lih (13:10.678)
Mm-hmm.

Deb Howes Studio Com (13:17.522)
I mean, I hope we're in a place where we can be better collaborators, but the challenge of getting not just museums, but like universities, museums, technology, whatever we need to make this happen, it just seems overwhelming. And I don't know, do you have a different perspective or something to lift my cloud?

Andrew Lih (13:40.697)
I'm out.

Deb Howes Studio Com (13:45.202)
I really want this to be the siren call for museums to understand that this is a huge opportunity for them to become more relevant. But this is not obvious to anyone, practically, except for you and I and a few other people who have really been in the field for a while. And this is a barrier that will get thrown back at us.

Like, oh my God, collaborating with all those people, like who has time, who has that money? Who, you know, I don't know. Do you have any clarity to bring to this? You may not, it's okay.

Challenge of collaborating at global scale: wikimedia to rescue.

Andrew Lih (14:20.778)
No, I think your sentiment is there and we see it all the time. In recent years, there's been something that's very similar called the American Art Collaborative. And that was something that predated kind of what we see now with Wikidata and a lot of the collaboration happens there.

The not the only bright spot, but one of the bright spots of a lot of this activity now happening on a platform like Wikipedia, Wikidata and Wikimedia Commons is that it provides a very inexpensive

pretty informed community, I'm sorry, an inexpensive platform, let me rephrase that. So the nice thing about all this collaboration and happening these days on such a popular platform like Wikipedia and it's what we often point to as two partner projects, Wikimedia Commons where the multimedia is hosted and the Wikidata, which are where the structured data is, is held, is that it is so well known, it is so widely regarded and so widely used, right? Because Wikipedia and its associate websites are

among the top 10 most visited web properties in the world. And it's free, which means that there's a much higher chance of being used. It's not just in little enclave or one little.org site. And it's multilingual, right? So you actually have a lot of value add because once you put up an artwork or metadata, you have other people helping to enrich it and to make it available in other languages. The saving grace about this is that efforts hopefully put into contributing.

artworks in this multimedia and data commons that is in Wikipedia have high impact and have high visibility. That's the hope. Because we all know that there have been these little spurts, whether it's art store or American Art Collaborative of maybe 10, 12, 15 top museums banding together to do something. But where's the visibility in that? Where does it go? How do you get publicity for that? And that has always been a big missing piece of the puzzle. It's nice that

Deb Howes Studio Com (16:00.591)
Yeah.

Deb Howes Studio Com (16:11.238)
Right.

Andrew Lih (16:17.578)
Maybe the Met and these big museums that are well resources combined forces, but where's the impact going to be? And the nice thing about today is impact is you upload to Wikipedia or Wikimedia Commons and boom, it's ingested into Google within 10 minutes. You say, hey, Google, who painted this painting? Oh, well, that data you contributed as a museum, it's now being announced on voice assistance. It's being available on Siri. It's made available on mobile phones. So there is that direct line between putting the effort in here from a.

Andrew Lih (16:47.274)
a cultural heritage institution to it being immediately useful, not like days or weeks later, but minutes. We've measured this before. We're uploading a Wikipedia article about an artwork or a person or something, makes it available within 10, 15 minutes on a Google search. And then by the end of the day, on Siri knowledge on every iPhone in the world, that's pretty amazing to think about that. And that has been a difference maker when we not only explain it, but show it to.

Deb Howes Studio Com (16:55.878)
Right. It's amazing.

Measuring impact

Andrew Lih (17:14.614)
museum executives and folks who are not quite convinced engaging with a site like Wikipedia is worth it. When we can show you that there's a direct connection between contributing here and showing up and having impact over there. Now measuring that and determining what that impact is, is a little bit tougher, but at least that wow factor is there to show that there is impact, there is something that's immediately useful to the public. But trying to quantify what that impact is and the ROI, the return on investment,

for an institution to undertake projects to work with Wikipedia has taken a while. We're lucky that we actually have prominent cultural and heritage institutions now that believe in it. So I'm working with the Smithsonian as a Wikimedia at large to work across all its units on Wiki related efforts, because that is a direct association to their goal of reaching a billion people around the world with digital content from the Smithsonian. And that's especially relevant during COVID when all these museums

Deb Howes Studio Com (17:56.405)
Yeah.

Deb Howes Studio Com (18:09.458)
Yeah. Yeah.

Andrew Lih (18:13.462)
Doors are closed. And it said, we have no footsteps coming in through the front door of museums. How do we still have an impact? Well, fortunately, we have these projects with Wikipedia where you can contribute the digital content. And now it is available not only on the Smithsonian website, but also in Wikipedia and any other language editions of Wikipedia that are being used.

Deb Howes Studio Com (18:34.626)
It's awesome. Andrew, what you're doing is just awesome. Um, has, do you know if this recent advent of chat, GPT and other large language learning models that we're getting access to, Dali, for example. Um, do you know if they've had an impact on how the Wikimedia foundation sees the future of their activities, products, support collaborations, like

AI impact on Wikipedia?

Are there things we can look forward to that you can talk about that might change our thinking about how we museums could contribute?

Andrew Lih (19:13.002)

Yeah, good question. Definitely there's a lot of people in the Wikipedia community that are experimenting with chat GPT in different ways. The interesting thing is that Wikipedia has always pride itself on being very kind of artisanal and hands on. In other words, the beauty of Wikipedia is not any technology per se, but the fact that it is very much human beings connecting with each other to collaboratively write encyclopedia editions and articles and things like that. So.

So certainly something like ChatGPT has interesting implications on two sides. One, the paradigm we've had for the longest time where you ask Google, you just type in a few keywords into Google, and it generally spits out not the answer per se, but a list of websites that are relevant. But we know that's changed, right? Sometimes Google will put into a box all the relevant information, and you're like, oh, I don't have to click on any links. I actually can find out the birthday of a.

Charlize Theron by looking at this box. Or I can see what Joaquin Phoenix looks like by looking at that box without ever clicking on Wikipedia, even if that content did come from Wikipedia. But now there's a direct path from question to explained answer. And it might have used Wikipedia, but you never see the word Wikipedia come up, and you never visit Wikipedia. That is exciting, but also a massive risk as well, right? Because.

Risk of wikipedia becoming transparent

In the old days before chat GPT, by having your Google search result show a bunch of pages, and you're always seeing Wikipedia at the top of the list, Wikipedia, Wikipedia. Or even if you're given the answer in that what we call knowledge panel says, Charlize Theron, born this date, according to Wikipedia. You're like, OK, Wikipedia gets the branding. Wikipedia gets recognition. On a whim, you might click on the Wikipedia link, and you'll go to Wikipedia. And you might get involved with Wikipedia. You might learn more from Wikipedia. You might donate to Wikipedia. What if all that goes away?

Andrew Lih (21:03.414)
that the content from Wikipedia is being used, but you never hear the word Wikipedia. That is kind of a long term or even a short term risk for Wikipedia in terms of staying relevant, even as its content is being used massively by these AI systems. So being cut out of the loop in terms of branding, recognition, or referencing, which is a big problem people have with ChatGPT, is that it gives you an answer, but it doesn't show the work. It doesn't necessarily tell you exactly how it got it. It doesn't even exactly show you where.

Synthesizing probability vs truth

the sources are. In fact, one of the major criticisms of ChatGBT is it is basically synthesizing probabilistic content. It says, probably, if you ask me questions about this author, this is what is likely. I don't even say the word true. This is likely what you're bound to see. But if you ever ask a question to ChatGBT about something that's not true, then you're bound to see it.

or someone or something that doesn't really appear in Wikipedia, it just makes things up with no shame. It'll say, you ask like, who is Dr. Simon Finkelbuster? And it'll say, Dr. Simon Finkelbuster is a microbiologist at the University of Iowa and he was educated in Germany. It's like, this is a fictional person. You made up all that content. But it won't give you any indication it was made up. So that is one of the biggest risks of chapter GPT as well.

is that this point, and I think it's going to get better, and it already is getting better, it doesn't show its work, and it doesn't really make a distinction between synthesis, fiction, and fact. That's kind of bad for information accuracy.

Deb Howes Studio Com (22:42.762)
And is Wikipedia able to negotiate on that at all with these tools? Because frankly, I have an example that supports what you're saying. Like, of course, like everyone else, I've been playing around with Chat GPT and I was asking questions that are relevant to this panel. And I would say, please list your sources. When I would ask a question like, what is the role for AI and museum learning? And it came down actually with reasonable.

Andrew Lih (22:50.731)
Mm-hmm.

Andrew Lih (23:05.559)
Mm-hmm.

Deb Howes Studio Com (23:11.798)
you know, like sort of a 101 reasonable level of like information. I don't think it was inspirational, original, or that interesting, but it knew what I was talking about and it could respond. And then it lists sources. And the sources weren't, no, they were real sources, but two problems I saw never listed Wikipedia. Now, I didn't know, or Wikimedia. I don't know that.

Andrew Lih (23:24.110)
Mm-hmm. Were they real sources or made up sources?

AI sources

Deb Howes Studio Com (23:39.498)
I didn't really see if anything that I could recognize actually came from Wikimedia. But what you're saying is that is Wikimedia content so deeply ingrained in their learning, they don't see it as a particular source to credit. I mean, I think my superficial reaction was, wow, it's never crediting. I've asked it a lot of questions. It's never crediting.

Wikimedia anything. Second is, I think I was asking something very specific about AI and museum data or something like that. And it quoted an article from 2017 in museums in the web conference, which you know, they do publish scholarly articles. And I said, can you, can you use a more recent reference because

Andrew Lih (24:28.428)
Bye.

Deb Howes Studio Com (24:36.266)
In 2017, artificial intelligence wasn't available and it shut down. Like, so I don't know if that was just a busy moment and, you know, I got a free version. Maybe if I pay them, like I'll get more better answers. But, you know, that, that kind of logic, I think really, I got the spinny wheel. I don't know if that out.

Andrew Lih (24:56.778)
Right. Right, right. And in general, in the field of AI, that's what we generally call explainable AI, or lack of explainable AI. The magic of ChatGBT is most of the time, it gives you something that's pretty darn interesting and many times very accurate, or much better than you could do in 60 seconds of Googling. So there's wonderful things about ChatGBT that are really useful. On the other hand,

When AI gets it wrong

Andrew Lih (25:27.318)
Um, when it's wrong, it's disastrously or hilariously wrong. And if you can't tell the difference, then is chat GPT useful is kind of the existential question. And for the Wikipedia community right now, the answer to is chat GPT useful for us right now is not really because Wikipedia really prides itself on over the 20 some years. It started off as, Hey, isn't this kind of a cool project where you can find out pretty good knowledge, but it's not highly reliable.

For the last 20 years, people are relying on Wikipedia now. They're really making life and death decisions about it. You could make or break someone's career if there's accurate or inaccurate information about someone on there. So now the Wikipedia community has really turned to taking information reliability very seriously. Where is ChatGPT now to fit into that formula? And right now, I think the right decision has been made. Danger and caution. You can play with it, but we don't want any massive use of it right now. And I think that's smart.

Danger and caution using AI

Andrew Lih (26:25.106)
until you have ChatGPT show all of its work or enough work that you can dissect like, so Simpson Fingal Bottom went to University of Iowa, how do you know that? And as you said, if it goes, okay, I admit, I made it up. Then we're better off than it just going blank. It's like, I'm not explaining anything to you. Because right now, I'm not explaining anything to you or I have no way to explain it equals unusable on Wikipedia. And I think that's the right path for now.

But the technology is going to get better, so that it will show more of the work and show you what things you can snip and say, no, that's inaccurate. This is accurate. But unfortunately, the large language models in the neural net systems, in general, have this kind of magical output that you really cannot explain very well. It's just that in most of those applications where they're

synthesizing a picture or they're synthesizing something, is it a disaster if you're not able to do that?

the picture of that girl has a sixth finger, which is a very typical thing you have in generative images. Yeah, you know, you edit it out in Photoshop and you're OK. But that's not something you can do when you're generating Wikipedia articles or life and death articles about diseases and things like that Wikipedia. So we don't want any of that inaccuracy. We don't really want the inaccuracies of a chat GPT type system proliferating in Wikipedia. We want to understand a lot better. So I think right now it's a step carefully.

play with chat GPT, but it should have no role until it's more explainable and we can dissect and audit what it's spitting out.

Deb Howes Studio Com (28:01.350)
And just to close the loop, you feel that the more images, authoritative information, especially in a non-Western context, I'm just going to say that, but it means a lot of different things, not just non-Western. Broader context for the objects, more points of view represented in the data structured

Andrew Lih (28:16.305)
Bye.

Deb Howes Studio Com (28:30.622)
Wikipedia, do you feel that that is the most effective way for us to help chat GPT get better?

Andrew Lih (28:40.472)
Repeat the question again.

Deb Howes Studio Com (28:42.294)
We, so, um, following along your line of thinking, which is that chat GPT right now is a danger zone, unless you're fully aware or fully unless you're an expert, unless you're an expert, you can't really find that much use in chat GPT because you have the knowledge to know whether what it's saying or not is true. And, um, so, but we see a day where.

Andrew Lih (28:53.514)
It can explain its conclusions. Yeah.

Deb Howes Studio Com (29:10.598)
chat, GBT and other large language learning models are going to be learning more and more. And as the museum, this is a museum conference, we're talking about museum learning, is our role to create as much solid, wide point of view, well, you know, global perspective information into to add it both as images and text and other media formats.

into Wikimedia, is that our best way to make chat GPT go from foe to friend, essentially?

Museums can improve AI quality

Andrew Lih (29:49.066)
Yeah, I think more original source material, more unique holdings materials, whether the imagery metadata or anything that would definitely help fill in the sum of all human knowledge, which is the tagline in Wikipedia. We think we're doing a pretty good job in Wikipedia with six going on seven million articles. That's like 10 times larger than Britannica ever was. It's quite an achievement.

But it's still a tiny sliver of what probably should be in a encyclopedia of the world or of all of human existence. We're still missing things about geography and monuments and people and all types of things. So there's a lot of work to be done there. So making sure the data set is more complete and well described is going to make the AI that's based on that better in the long term. So that's something that.

Deb Howes Studio Com (30:28.378)
Yeah.

Andrew Lih (30:45.734)
is certainly something that cultural and heritage institutions can help contribute to.

Deb Howes Studio Com (30:52.278)
Well, the reason I'm drilling down on this is because a lot of not a lot but the major museums around the world are like I made an API my whole collection is out there Like why can't they why can't that just be you? Why do I have to put it into Wikipedia? What's your answer though?

Andrew Lih (30:58.165)
Hehehe

API problems

Andrew Lih (31:03.158)
Right. Right. Yeah, APIs are great in terms of having one institution's particular lens on something. Like the Met has an API. The Art Institute of Chicago has an API. Getty has an API. Smithsonian's API. Those are great in terms of having mechanistic or programmatic access to things at scale. But they are not all interoperable. So each API is unique in terms of how it works.

Andrew Lih (31:32.898)
They'll answer to certain types of questions, but they won't answer necessarily the same questions across APIs. So certain APIs are optimized for searching on artists or on artwork, but they may not be optimized to search in genres or in dynasties from Egyptian history or dates, for example. So one benefit of having it in a common corpus like Wikipedia or Wikidata or Wikimedia Commons for the images

Andrew Lih (32:02.226)
a common model, or sometimes, you know, we call it a common taxonomy to describe things, right. So it's not. It's not different ways of describing things that pass each other the night, but they're all, they're all kind of correlated. You know, a really simple example that we run into all the time is, what is a painting or what is a picture, right? So for example, if you have a leaf, right, which is a page out of book,

Some museums see that single page as a painting, and it's classified as a painting. But some other museums might not. They say it's a leaf of a book, and they classify it as a leaf. So if you're trying to count how many paintings exist in these three museums, it depends, because some museums call it a painting, some museums don't call it a painting. Are things classified as sculptures or busts or things like that? So there's all kinds of different ways of modeling art. And there's no one specific answer sometimes. There are all different ways of describing artworks that may be.

Wikipedia is an unified ontology

different depending on if you're a museum oriented toward a certain type of style or different type of classification. So these are all things that are hard to resolve with different APIs and different models. But having it in one, at least with one kind of unified ontology that you find on Wikipedia or Wikidata is useful in having some kind of unified view of the world. And that's one reason why efforts like you described or.

American Art Collaborative are okay for a certain domain, but they don't necessarily connect to the entire world of artworks.

Deb Howes Studio Com (33:34.522)
Thank you. I think that really clarified some things for me. Um, I appreciate that. It made me also think about whether what's your thinking about. Let's just say the probably Silicon Valley or IBM, whatever, like the people who are developing these large language models, how interested in they are they sorry, in how interested.

are they in working with the cultural sector to make their training more globally aware, accurate, et cetera? What's your feeling about that, if any?

Is silicon valley interested in partnering with Museums on making AI better?

Andrew Lih (34:20.702)
I think there could be more resources put into it. One. Both. Well, there was one bright spot. We actually had a hackathon across four different domains back in 2018. So this was a weekend hackathon that was organized with the Met Museum, Microsoft, MIT, and the Wikimedia community. So I was part of this.

Deb Howes Studio Com (34:25.710)
By whom? By the museums? By the companies?

Andrew Lih (34:48.406)
this weekend, we were up in MIT in Cambridge, Massachusetts. And we put our heads together, like, how could AI be applied to this domain of open access content? The Met had just announced that they were releasing all the images and metadata under a Creative Commons zero, which basically means like equivalent to public domain. Anyone can use it for whatever they want. So what might the implications be for AI? So we actually had a weekend where we created

AI systems to generate new art based on the art from the Met. You know, back then it was a new thing called a GAN, a Generative Adversarial Network. Today, you know, we have probably dozens of tools for free on the internet that do this, but back in 2018 it was kind of cool to have this funky tool that generated new artworks based on existing artworks. We also had another project that tried to recognize features in artworks based on

training it on met artworks and what it had already seen as labels for those artworks. So I think, you know, that was just one example of a hackathon. I hope there are more of those coming down the pike. We're going to have our global Wikipedia or Wikimedia Community Conference in Singapore in August. And we hope to engage tech companies like Google and Microsoft and other folks there for that, you know, week in.

August to see how we might put our heads together and have, you know, as you said, the Silicon Valley type companies put more resources into working with Wikimedia content in a responsible way. But yeah, more effort could be done.

Deb Howes Studio Com (36:28.602)
Why Singapore? I'm just really curious like why Singapore? Is it with NTU or is it like the university there? Or their main event?

Andrew Lih (36:35.150)
Good question. So the Wikimania conference that we have every year tries to move to different places around the world. So we engage different constituents. So typically, it has been in, you know, European and North American countries, but we try to move to places in South America, Africa and other places. It's actually been in Alexandria, Egypt and South Africa before, but it hasn't been back to Asia in a while. It was in Taiwan and Hong Kong before. So this is this year in Singapore, which is

Quite interesting because even over the past 20 years of history of Wikipedia, Singapore has changed quite a bit to become quite a significant technology hub. So a lot of the headquarters for these tech companies in the Asia region are based out of Singapore. So those are pretty significant.

Deb Howes Studio Com (37:18.778)

And the universities have really amped up their programs in digital art making and kind of creative, the creative aspects of digital. So that's why I thought maybe you were being invited for that reason.

Andrew Lih (37:19.917)
Yes.

Andrew Lih (37:30.038)
That too, that's something we hope to engage as well. You're right that both Hong Kong and Singapore and increasingly China as well, have done a lot to do with computer graphics, special effects and visual arts. And that's something that could be a great thing too.

Deb Howes Studio Com (37:49.554)
Before I let you go, I just want to have a moment to focus on the fact that this conference is about, it's a museum learning summit. And as a museum educator, I always try to come at a problem, a question, whatever, from the visitor point of view. And one of the things that as someone who's spent a lot of time in the galleries talking

with people in my gallery talks or just randomly, people who would say, describe themselves as someone who knows nothing about art, for example. Usually I'm working at an art museum. But then they explain what they like or what they're looking for or what their interests are. And I realized that

Deb Howes Studio Com (38:41.810)
Most of the museums I've worked in are encyclopedic where they organize their collections in some kind of chronological sequence, which as an art historian, I totally get. Like it's not a problem for me to find exactly what I'm looking for. But when the public comes in and they're using natural language, not sophisticated art scholar terms to describe to someone like me, who they see as a translator, to direct them to what

they're looking for. I feel like that is what you said the magic of chat GPT is that that's kind of the promise right the promise of chat GPT is that I can use my words that I'm comfortable with to describe an outcome that I want and it will match and what you've already identified is that it's not matching and they don't know.

Andrew Lih (39:26.619)
Mm.

Andrew Lih (39:29.027)
Right, right.

Deb Howes Studio Com (39:39.098)
that it's not matching unless they just look at it and say, no, that's not what I meant. And then they try something else. But if they're actually looking for words, they may not know whether that

information is accurate or not. So there's a little bit of a disconnect based on the visual. So that's just a little background by saying, what I always found to be stressful about museum learning in a digital space like the internet is that it's not visual enough.

Andrew Lih (39:40.290)
Mm-hmm. Right, right.

Andrew Lih (39:49.367)
Right.

Deb Howes Studio Com (40:08.750)
Like if people could just let go of wanting to find this one certain thing, which a lot of people do, like they go to an art museum, they're like, whatever the Met wants to show me, whatever the Met thinks, or the MoMAs thinks is important, let me just walk around and I'll say yes, no, you know, and sort of wander around. That kind of learning, there may be some, like a name for that kind of learning, but that kind of let me wander around, see what I think.

Andrew Lih (40:12.247)
Mm-hmm.

Deb Howes Studio Com (40:39.762)
and follow paths that I like is, I think, not only missing from chat, it's also missing from Wikipedia because I don't always know the right words. And of course they have elaborate like referrals and whatever. But the visualizing, let's say of chronology or the visualizing of global cartography or the visualizing of books.

Andrew Lih (41:00.287)
Mm-hmm.

Deb Howes Studio Com (41:07.810)
that you were mentioning, well, is it a painting? Like, I wouldn't even think, you know, if I was looking for one or the other, it wouldn't even occur to me, like you said, that they cross over. But if I saw images of the illustrations, you know, visualized to me, I don't know. Like, I was hoping that Dali would sort of go in that direction. But in my little playing around with it, it really isn't very knowledgeable about what it's showing you.

Andrew Lih (41:27.138)
Hehehe

Deb Howes Studio Com (41:37.618)
And I think it's not very satisfying. Like when you go to a museum, you know there's some kind of logic, which has a certain kind of veracity, let's say. It may not be my truth, but somebody's truth said these things are important and this organization of them is interesting. And what do

you think? Open the gallery, what do you think? All right, so how can we bring that kind of learning?

Andrew Lih (42:01.847)
No, that's-

Deb Howes Studio Com (42:07.722)
into this AI space. I really don't know.

Lack of visualizations in Wikipedia and internet in general.

Andrew Lih (42:10.070)
Yeah, that's a good question. And I share the frustration that after working with Wikipedia for 20 years, I've concluded, and I've written a book called The Wikipedia Revolution. I still believe it is a revolution on how we assemble knowledge and how we participate in writing history together, right? But I've also concluded that it's a limited revolution and that it's a revolution in how to create a very conventional encyclopedia.

Deb Howes Studio Com (42:21.818)
Yeah, I know that book.

Andrew Lih (42:36.106)
Because Wikipedia is still, as you noted, a static set of pages that actually have policies against too many images. Its primary policies in Wikipedia around images is not to put too many images in because there is still this, without being too insulting, this still very provincial idea that serious encyclopedias are text.

Children's textbooks have pictures. The more pictures you put in there, the less seriously they'll take us. Therefore, we need to stay text-based. There is still that bias in Wikipedia. There are pages in Wikipedia saying, do not include too many images. Only use what you need to. When's the last time you saw a video in Wikipedia?

Andrew Lih (43:27.678)
I can't point to a single video I've seen Wikipedia for the last three, four weeks. You go to the article about Tango, about salsa in Wikipedia. What do you see? Text and a picture. We should be watching video. We should be seeing animations in front of us. You go to an article about missile launch, parabola, things like that. They're not animations. They're not simulations. They're not interactive. They're just text and pictures. So Wikipedia is still very retrograde when it comes to.

Um, being multimedia rich to educate us. So I think you hit on something exactly as you should have hit on. It's like, so when's this next, what can either chat GPTU, any kind of revolution of, um, platform do for enhancing the experience of open knowledge to folks. Right. So my favorite

example of this potential is, and I can show you the, the, the visual later on, if you want to slice it in.

Mapping in Wikidata

My favorite example of this is what you can generate out of Wikidata, which is the latest project we have in the Wikimedia sphere. It's what happens when you structure all the content we have in Wikipedia. So you read the Wikipedia article about the US Congress and says, United States Congress is a bicameral legislature of the United States. But what if instead of reading a sentence, you could break down that sentence into many factoids, which is basically the atoms of information that we have in Wikipedia?

Congress is bicameral legislature. And we know that bicameral legislature is made up of two houses, right? And we know a house of a legislative body is X. So you start to see all these connections, right? But what if we could visualize all those connections? And once we visualize those connections, we can start to see like, well, what other bicameral legislatures are out there? But how does India's works different than the United States one? And what's the history of how they did? So you can start to see these connections through Wiki data this way.

And my favorite one is ones related to museums. And I think there's huge potential for museums to do this. And we just haven't found a repeatable way to do this. But my favorite one, the Met, is the portrait of Madame X. So some of you may know this. But the weird thing about the Met, as a weird sidebar, the weird thing about the Met Museum, which I love, obviously, because I work with them, is that there is no kind of seminal artwork that you associate with the Met, which is weird, right? The Louvre, you say, oh, Mona Lisa, of course. You say British Museum. Oh, the Rosetta Stone.

Deb Howes Studio Com (45:53.874)
At the Art Institute, there's Edward Hopper Nighthawks, there's American Gothic, yeah. That actually was a big shock for me when I went to the Met from the Art Institute. It's like, what's the signature image?

Andrew Lih (45:56.134)
Exactly. Right. Right. You have like the iconic artwork.

Andrew Lih (46:02.658)
Is that right? Right. Right. Well, you could see Frank Lloyd Wright's house. We have his room from Frank Lloyd Wright. We've got Portrait of Madame X. We also have the Death of Socrates, if you want to see that. And it's like, oh my god. It's not that you don't have one. You have so many that they're all kind of iconic. But no one stands out there. But a favorite of a lot of people is the Portrait of Madame X. John Singer Sargent. Huge piece. Really beautiful.

Andrew Lih (46:31.990)
kind of posed by this woman. And there's a lot of interesting stories behind this painting because it caused such a sensation in Paris. It was revealed that he had to flee the country. There was such a controversy around how scandalous this painting was. There's so many stories in this

and so many connections that it'd be great to experience this not just as text on the page, but the seeing all these connections. So if you actually go into Wikidata, where we've actually modeled.

information around this painting, like who painted this thing? Where is it hanging today? Who is this woman in the painting? What other paintings are there of this woman? What was the impact of this painting? Oh, you know what? There's three books written about the painting. Not three books written about the author. This one painting has three different books about it because it's so scandalous. Oh, and by the way, you know that scandalous dress she wore? That was actually the basis for this dress that was worn in the movie Gilda by Rita Hayworth. And you're like, wow, I didn't know that.

But it's a lot more interesting seeing this in graph form to see all these things. And the reason why you can do this, and it's not just someone decided to write about it. You can actually discover these connections in Wikidata is because Wikidata is like the database analog of Wikipedia. Wikipedia has this tagline, it's a sum of all human knowledge in text form. Wikidata is like the sum of all databases in the world, but in like graphical form.

And the beauty of Wikidata is it's not just a arts database that you might be interested in as a Met person. It's also a fashion database. It's a literature database. It's a film database. It's every kind of database. So the cool thing is by starting with the painting, Portrait of Madam X by John Singer Sargent at the Met, which are like three nodes on this graph, you can start to see all these connections that say, oh, you know what? That painting inspired this designer to make that dress in this movie. And this person wrote this book about that woman in that painting.

And suddenly, it's visual, it's rich, it's exploratory. It is something that allows you to discover things you never knew. And we're still discovering things we didn't know about either that painting or other things out there. So I think that's the potential we have out there. And what if we could now, long way of saying, insert AI and ChatGBT into that equation, to say, ChatGBT, tell me something about this painting that is not that well known. Or help me. What?

ways has this painting impacted cinema and have it help explore and find those connections. I think that's a huge potential for AI is to probe, explore, test these different connections to assist us. And I think that's the main thing that we've concluded at this point in time in 2023 with ChatGBT. It's a very useful assistant. I would not rely on it for the final word on anything. I would not rely on it to be the ground truth of anything.

But to kind of get you through a lot of the hard work to comb through things, to associate things, to find the activation energy to write something, it's actually been very useful in that area. So even I know I bad mouthed chat GPT before about, can it generate a Wikipedia article? No. But it's not a bad place to start for drafting something, as long as you fact check every line that it spits out. Because it has been useful for some folks in community to say, hey, we don't have an article about this new

Andrew Lih (49:49.002)
winner of this prize. Get ChatGPT to start drafting the Wikipedia article. And in most cases, it does about 60, 70% excellent job. But you need to fact check everything and source them.

Deb Howes Studio Com (50:00.278)
Yeah. You know, you just reminded me, I haven't actually tried this, but I'm going to right after we're done. I haven't actually restrained chat GBT to say use only Wikipedia to answer this question. And I wondered if it would do that. That's so interesting. I am reluctantly going to let you go because I don't want to take up more of your time. You have taught me so much in this less than

Andrew Lih (50:12.502)
Yeah.

Andrew Lih (50:17.495)
Yeah.

Deb Howes Studio Com (50:30.074)
And this video is going to be awesome because you've been so generous and I love, it's not, I've always gone to your sessions and you're always telling me what's up to date, but this is the first time I got to hear Andrew Lee talk about the future. And it's just super exciting. Yeah, I think I have a link from one of your slides that you gave me. I have a link to this.

Andrew Lih (50:43.949)
Yeah.

Andrew Lih (50:48.471)
Bye.

Andrew Lih (50:54.646)
Yeah, you probably have a link to this. Yeah, so you can show this if you want, but that's the painting I was talking about there, obviously, so.

Deb Howes Studio Com (50:59.809)
Um.

Deb Howes Studio Com (51:02.458)
Fantastic. Yeah, we're going to have some ability to show visuals. They just can't be slide. Um, and, uh, I love this story. My, you know, that the project that I'm most known for at the Met was building the timeline of art history. And this was of course, in a time when even the Met didn't have one database for all of their collection. So there was very little automation that we had, but something that

Andrew Lih (51:16.351)

Yes, that's great.

Andrew Lih (51:22.924)
Mm.

Deb Howes Studio Com (51:29.526)
We've now lost and I should write an article about this, but you know, as, as the Mets website has evolved through different platforms, it's lost this visualness that we of course handcrafted. And when you were talking about the navigating from the Wiki data, I'm like, there's gotta be a way to generate a visual visit from that Wiki data. Maybe we don't have enough information right now.

Andrew Lih (51:38.646)
Right? Yeah.

Deb Howes Studio Com (51:57.146)
But if we did have an AI tool that instead of, you know, showing it like a graph like you did, but would show the connections visually, which is what we tried to do in the timeline in the old model. I don't know if you can see it behind me, but it's like all these faces line up, you know, like that was what we did is we, we by hand made a visual timeline of things that were happening, let's say in, you know, Baghdad.

Andrew Lih (52:14.211)
Mm-hmm.

Deb Howes Studio Com (52:27.190)
in this particular moment in time. And that would be awesome. If we could, you know, I feel like for museums, this is what I hope, that we can coordinate like you did with MIT, maybe with IBM, maybe with Microsoft, to work together to kind of visualize the history of the world as seen by artists, that would be awesome. I think people would really love it.

Andrew Lih (52:46.678)
Mm-hmm.

Andrew Lih (52:50.750)
Yeah, I don't know if I will show you that we actually are recording into Wikidata, whether it's in the timeline of art history right there.

Deb Howes Studio Com (53:00.819)
Um, Oh, you mean, I'm sorry. Tell me what I'm looking at again.

Andrew Lih (53:02.742)

This is the Wikidata record for the Portrait of Madam X. So I was saying that all your work is not being lost. We are actually recording into Wikipedia and Wikidata, whether it's an entry in TOA. We call it TOA. I don't know if it was called TOA back then. Yeah. Yeah. Good.

Deb Howes Studio Com (53:07.957)
Oh, okay.

Deb Howes Studio Com (53:17.814)
Yes, no, that started a long time ago. I'm familiar with that moniker, Toa. Yeah, I'm glad. No, I don't think that anything's lost. It was just in terms of the visual interface, right? That now we always, when the visitor goes to the Toa, they have to choose like something different than from a map or from a visual interface. So anyway, but I feel like that visualness of that exploration.

could be a really interesting thing for my AI assistant, like you're saying, to guide me through.

Andrew Lih (53:50.135)
Mm-hmm.

Andrew Lih (53:53.438)
I think that's a great vision for that because...

Andrew Lih (54:00.758)
And I'm glad I had the conversation with you, because I think that is a really interesting project to try out in a future hackathon, is to say, what are the strengths and weaknesses of using something like ChatGP to find those connections that we don't yet know of, right? And to steer ChatGP in the right direction. Right. Yeah. Very cool.

Deb Howes Studio Com (54:10.478)
Let's make it happen!

Deb Howes Studio Com (54:21.318)
Love it. Okay. Andrew, thank you for everything you do. You are definitely making this world a better place, at least in the cultural sector. I can say with authority. And, you know, I just really appreciate you. I just want you to know that.

Andrew Lih (54:28.705)
Thanks.

Andrew Lih (54:35.931)
Oh, thanks so much. Great talking to you.

Sophia (54:39.766)
But here we go. Sorry. Thanks. That was really cool. Okay, guys, I'm excited to share it with you. And I just wanted to say we didn't get an intro per se to this. So if you wanted to say start off,

you know, by saying what this talk is and introducing Andrew or Andrew if you want to introduce yourself, we could.

Deb Howes Studio Com (54:41.712)
We can go ahead so...

Deb Howes Studio Com (55:04.642)
You know, I think what I'll do, because the talk is 20 minutes. And so I think what I'll do is I'll record an intro with Andrew off camera. Uh, because I haven't written it yet and I want to do a nice job. Um, and, uh, so it's really just to accommodate me and my, um, behind the schedule situation, but it's great that you thought about this. What I want to say to you, Andrew, before I'm sure you have to pop off is.

Sophia (55:07.509)
Yeah.

Sophia (55:15.390)
Okay. Cool. You're not, what are you?

Sophia (55:24.422)
Great, now that's totally fine.

Deb Howes Studio Com (55:34.462)
We're going to generate a transcript from this. I'm going to send you the transcript. If there's anything you want us to take out, just go ahead and let us know what that is. We can also show you the video in this raw state, but what I'd rather do just to save time on your side is that once I get the transcript back from you, or you could just say it's all fine, whatever you want to say, we're going to, as you know, we're going to record Douglas on Friday.

Andrew Lih (55:51.425)
Mm-hmm.

Deb Howes Studio Com (56:01.826)
I'm gonna come up with a kind of an overall script. I'm not yet sure whether I'm gonna create questions and answers for the two of you to respond to or whether it'll just be sequential, whether I talk with you and then I talk with Douglas. This is something that I'd like to spend some time in once we've done the recordings, which is next week is my sort of think time. But then I'm gonna show you some near final draft by the end of this month.

And then I have to submit this, like, I don't know, some absurd amount of time before the actual date on the 17th of June. I mean, it's just so much work for Museum Next, but I'm so supportive of them having a digital learning summit because, you know, we often get educators, museum educators in particular, we often get sidelined for different conversations. And so I really want to support the focus of what.

they've done here at MuseumNext. Any questions?

Andrew Lih (57:01.586)
Yeah, it's now I don't know. I mean, the funny thing is, I gave a talk at New Museum Next many years ago when it was at Bloomberg, if you remember those days or but I don't really know much about them. I'm glad they're still here. So that's great.

Deb Howes Studio Com (57:13.091)
Yeah, I do remember that one.

Deb Howes Studio Com (57:19.118)
Well, they did a lot of transformation during COVID, no surprise. And I think most of what they do right now is virtual, where it used to be, they would, like when I was at MoMA, they sort of took over MoMA for a day and a half. And we had all the auditorium for them and they had the speakers going in and out like we used to do before COVID. Um, and I'm sure that was like a big expense that then they had to, you know, figure out.

Andrew Lih (57:21.725)
Mm-hmm.

Andrew Lih (57:25.633)
Uh-huh.

Andrew Lih (57:31.992)
Mm-hmm.

Deb Howes Studio Com (57:47.534)
like with the vendors and get advertisement. But their audience is a little different than like AAM, which I'm also going to in two weeks, or Museums in the Web or MCN. Their audience are CEOs, museum directors, a lot of people in promotion, and they don't really focus on education at all. So the fact that they're having this digital learning summit.

I think is a good thing. I want to support it. I'm hoping that some of the people who usually go to MuseumNext will come to the Digital Learning Summit. I really haven't talked to the organizers to see how they're targeting it, but they do have a Digital Museum Summit just before that. I think it's actually this month, might be the end of this month. I don't know if I'm gonna go or not, but that is something they've done before.

Andrew Lih (58:27.392)
Mm.

Deb Howes Studio Com (58:47.446)

So the fact that they have that and then this digital learning summit only a week or two in between is sort of interesting. Also important to know they're out of the UK.

Andrew Lih (58:52.653)
Yeah.

Deb Howes Studio Com (58:58.903)
So it's a better international audience than what some of our usual US bound ones are, like MCN and Museums in the Web. It's definitely.

Andrew Lih (59:05.729)
Right.

Andrew Lih (59:08.398)
Right, right, that makes sense. Good. Well, it's been fun.

Deb Howes Studio Com (59:16.266)
I'm glad you enjoyed it too. Um, okay. So you'll, you'll get like annoying emails from us. Um, hopefully.

Topics covered with Andrew Lih 4-26-23:

- Compares today's general understanding of AI with mid-1990's discovery of Internet/WWW. "we are on the verge of another information revolution on this scale: everyone needs to harness the power of AI"
- Explains why Wikipedia/media is at the core of the AI universe:
    - open online data systems vs. closed (e.g., online collections)
- References data bias (towards Western Civ) of much cultural information and steps taken to rectify this
    - metadata standards such as Getty indexes, library cataloging... need exapansion
- Mentions the difficulty of museums in collaborating towards uncertain ends (e.g., AMICO, Linked Open Data)
    - e.g., American Art Collaborative on cataloging data from multiple perspectives,
    - these efforts have not been as scalable, accessible, or as sustainable as Wikipedia/data
- AI poses a new risk to Wikipedia because WP sources are not revealed automatically--becoming like an absent partner with no visibility.
- Discusses another major criticisms of ChatGBT is that it is basically synthesizing probabilistic content/not Truth
    - " it doesn't show its work, and it doesn't really make a distinction between synthesis, fiction, and fact. That's kind of bad for information accuracy."

    - " But unfortunately, the large language models in the neural net systems, in general, have this kind of magical output that you really cannot explain very well."
- Encourages museum to focus on making Wikipedia/media the best it can be:
    - museums are critical to "making sure the data set is more complete and well described is going to make the AI that's based on that better in the long term."
- AI's problem with API's: "they are not all interoperable. They won't answer necessarily the same questions across APIs. So certain APIs are optimized for searching on artists or on artwork, but they may not be optimized to search in genres or in dynasties from Egyptian history or dates, for example."
    - also not flexible around terminology, i.e., connecting natural language with scholarly words
- Showed optimism around collaborations between cultural sector and AI research/buisness operations.
- Limitation of Wikipedia on visualizing information:
    - "it's a revolution in how to create a very conventional encyclopedia. Because Wikipedia is still, as you noted, a static set of pages that actually have policies against too many images. Its primary policies in Wikipedia around images is not to put too many images in because there is still this, without being too insulting, this still very provincial idea that serious encyclopedias are text."
    - shows the data visualization of Madame X in Wikidata as a start in this direction
    - considers the potential of something like TOAH to drive AI and serve as an online gallery guide for visitors.