FINISHED FILE

APRIGF 2026

OCTOBER 5, 2026
GOVERNING AI TRAINING DATA IN THE ASIA-PACIFIC: TRANSPARENCY, VERIFIABILITY, AND CREATOR RIGHTS
KANCHENJUNGA HALL
13:15 P.M. NPT

Services provided by:
Caption First, Inc.
P.O. Box 3066
Monument, CO 80132
www.captionfirst.com

This text, document, or file is based on live transcription. Communication Access Realtime Translation (CART), captioning, and/or live transcription are provided in order to facilitate communication accessibility and may not be a totally verbatim record of the proceedings. This text, document, or file is not to be distributed or used in any way that may violate copyright law.

Governing AI Training Data in the Asia-Pacific.

>> SODAM KUM: Hello, everyone. Works? Okay, thank you.

So hello everyone, again. Good afternoon, Ladies and Gentlemen. Thank you for joining us today.

It is a pleasure for us to present our work at APrIGF. So my name is Sodam KUM and joining me is Seungwoo Jang and Eusun Lee. Seoyeon HAN is joining us virtually from Korea.

We are the expert group on get governance organized by KISA. We focused on research.

And today we're expanding our research on Asia-Pacific context, so let's begin.

Before we start, let me briefly outline today's session. I'll start with on our research background and our key questions. Next, Seungwoo Jang will examine training data.

Seoyeon HAN will compare the policies of Korea, Japan, and Nepal. After that we'll discuss legal and technical conclusions. Eusun Lee will present our AIPAC governance proposal and conclusions.

Finally, we'll open the floor and look forward to hearing your perspectives and questions.

So across the Asia-Pacific, governments all building their own foundation models and multilingual datasets to strengthen the national AI capabilities.

Yet, creators often cannot tell whether their work was collected or used to train these models. Existing frameworks including the Berne Convention and WIPO treaties provide international copyright protection.

However, they were not designed to address the scale and complexity of AI training. The gap we focus on is, therefore, the lack of shared regional standard connecting three things. Disclosure of training data and whether people were able to exercise their rights.

We have two different questions among legal and creators.

However, for creators to exercise their rights, three stages need to be connected.

Disclosure, verification, and exercise of rights. The problem is that these stages are often disconnected. This makes training data a governance issue. Who controls the information, who can verify it, and how stakeholders can share responsibility for connecting these steps -- stages.

This brings us to our central question. How can we -- how can the Asia-Pacific reconcile national ambitions for sovereign AI with accountable governance of training data across borders?

We approached this through three subquestions. First, disclosure and verification. What information should developers disclose about their training data and how can creators check whether their own works were included?

Second, rights across borders. When countries have different laws and AI policies, how can creators refuse use, request [?] or seek compensation and pursue legal remedies?

Third, shared principles. Beyond national legislation, what rights should be considered?

Let's start with the practical question behind all of these. Can creators find out whether their work was used for AI training? The Books3 case helps us examine that problem.

>> Seungwoo Jang: My name is Seungwoo Jang and I will explain about Books3.

Okay, let me start to read the real case called the book story. Book story is a dataset made up of nearly 200,000 books. These books were collected with permission.

They were all copied from a public site listing in 2020. This dataset then became part of a large public dataset that many procedures to train AI.

Later [?] to train model like [?]

In 2023, it was taken down from its original website, but by then, copies had already spread so it did not really disappear.

Most of the authors had no idea their books were inside this dataset. But how did they find out not even a law or not even an office [?]. So it is just because when journalist built a search tool.

That is, governance to protect creators' rights are not fully built.

So after Books3 was released, something strange happened. For example, Nvidia was open about this data and its model card clearly listed the large public dataset that included Books3.

There are others used the valuable information to show Nvidia. Nvidia argued that one model did not use Books3, but the case moved forward.

Apple was in a similar position. It listed its datasets and the authors claimed that they contained books 3.

However, OpenAI was different. It didn't reveal dataset. During a lawsuit what we only learned was it had deleted them in 2022 before the case began.

So the companies that were open got sued first.

The companies [?] say less. That's exactly what we see. Meta named books 3 in its first LLaMA paper, but not in later models. According to the [?] transparency index which scores measure AI companies every year, the average transparency score fell from 58% to 40% in just one year.

This is not only American problem. Others found copies in books 3 and so [?] the harm is global and but not the way authors get paid in the U.S. system.

Take Anthropic settlement $1.5 billion in books. Only books registered with the U.S. copyright office could be paid.

Many books were never registered in the U.S., so they were left out. This is strange. So because under the [?] commission, copyright is protected automatically with no registration.

So there are three applications. First, verification depends on chance. Not institutions.

And second is voluntary transparency erodes, so we need a disclosure duties that apply to every in the independent verification.

Third, data crosses borders, but remedies do not. So APAC comments on principles.

Next, Seoyeon HAN compares Korea, Japan, and Nepal.

>> SEOYEON HAN: Thank you. I will compare these gaps. Let me start with Korea. In Korea, AI training's mainly under Article 35-5 of the copyright act [?]. In February 2026, the Korea Copyright Commission shared guidance and failures in generative AI training.

However, a focus is on whether training may be lawful, not whether specific work was actually used.

Korea's basic AI Act Title 4 for transparency. Under Article 34, providers of high-impact AI must explain an overview of the training data used.

When an overview is not released to a specific works, creators do not have a work [?] verification procedure.

Now Japan. Article 30-4 of the copyright act uses copyrighted work for [?] purposes, including AI training.

However there is not blanket extension. It does not apply when used unjustly, the interest of copyright owners.

Japan therefore has a [?] on when copyrighted works may be used, but the work procedure for creator to [?] works are included in the training guide as such.

Finally, Nepal. Nepal adopted its national artificial intelligence policy 2,082 in 2025.

The policy identifies data governance and intellectual property protection as important areas, but [?] disclosures and verification procedure have not yet been defined.

So these three countries are at different stage of development, but they have a similar gap.

The comparison shows an important distinction between permission and actual use. The Korea relies mainly on barriers. Japan has criteria under art 30-4. Nepal is still developing its AI and intellectual property framework. But these systems do not fully answer whether particular work was actually used in training.

In Korea, there's no work-level verification mechanism.

Japan does not provide a work-level disclosure procedure.

In Nepal, verification procedure has not yet been defined.

This brings us back to books 3. Also we were able to check whether the books were included, because the dataset become accessible, not because the [?] guarantee of verification rights.

This shows where the gap remains. The role can tell us whether the work can be used but it doesn't tell us whether their work was actually used.

Without that information, it is difficult for Korea to exercise their rights. [?] cross-border while creating rights and legal remedies still prevent many at national level.

For this reason, work-level verification becomes a broader governance issue. The next step is to consider whether existing legal and technical [?] can close this gap.

Eusun Lee the discuss that issue.

Thank you.

>> EUSUN LEE: Let's first look at the legal tools available for creators. There are three main opportunities, litigation, disclosure of duties, and opt out mechanisms.

Litigation allows user to challenge the use of their work, however, outcomes vary across countries and cases and creators need evidence to support their claims.

Disclosure duties for developers to provide information about their training data.

Opt-out mechanisms allow them to say they don't want their work to be used for training. But the standards can be unclear and opting out does not do past training.

These tools are useful but a central gap remains. Verifying whether a specific work was actually used. Can technical tools fill this gap?

There are some possibilities, but each as limitations. First, membership interference estimates whether a particular text was used for training by examining the models' responses.

However, it's reliability --

(Lost audio).

Restrictions and track data -- show that the developer respected it.

Finally, machine unlearning aims to remove the influence of particular training data. Complete removal, however, remains difficult to guarantee and verify.

Our takeaway is that these tools can support verification, but cannot establish reliable proof in every case. We, therefore, need governance procedures that connect technical evidence with accountability and creators' rights.

We can think of this process in three stages. Disclosure, verification, and the exercise of rights.

First, developers disclose information about their training data. Second, creators need to verify whether their own work was included.

Third, they need that information to support a claim or assess whether their requests have been respected.

Disclosure is improving, but it often remains at a level of websites or datasets.

Creators, however, need answers about individual works. There is also an information imbalance. Creators are asked to provide evidence while developers hold the relevant training records.

We therefore propose that independent parties should be authorized to examine nonpublic training data and verified use of specific works.

>> SODAM KUM: Putting the last three presentations together, disclosure is growing but verification is blocked almost everywhere and are so rights.

That raises the question who gives the authority to verify and how?

We see two complementary pathways, national reform which is slow but binding and a shared visual layer on top of existing treaties which is faster, but [?]

We propose doing them in phases. So what would that look like in practice?

We propose six principles, each with a concrete mechanism attached.

Transparency, original pilot for content-level disclosure for high-risk, high-impact models.

Verifiability and independent audit. Law should require technically Vice Mayor final viability and non-public data under confidential, public review.

It's not enough as the decline shows. Creator rights, collective compensation by work type or language community, as the U.S. Copyright Office discussed in 2025.

Interoperable opt out. IETF's AI preferences and TDMRep are still being written. We're proposing the region join the process now so multilingual and [?] get built in from the start.

And multi-stakeholder oversight. Extend Singapore and ASEAN soft-law models that they have already built through the WIPO conversation and I.T. technologists.

So why does this matter here now?

Legal development differs sharply across the region. Korea and Japan are refining standard while Nepal has only set the direction of its national AI policy.

Waiting for everyone to reach the same binding standard leaves the least equipped creators unprotected the longest. That goes directly to the point, safeguarding a resilient Internet in times of crisis.

The longer trust in training data stays unverified, the wider that gap grows and it grows fastest where the capacity to demand is weakest.

So phased regional approach lets every country move at its own speed without leaving its creators behind.

So returning to our research question, sovereign AI and accountable data governance are not opposing goals.

Actually, verifiable training data infrastructure is the foundation of a sovereign AI ecosystem that people can trust. Rather than each country reinventing the six principles alone, we propose building them as shared regional infrastructure.

We would now like to on the discussion.

>> These are the three main questions we'd like to discuss today. First is industry review. So please share your thinkings on the balance between work-level transparency and legitimate trade secrets.

And Number 2, it's about the sovereign AI, so please share your thinkings on whether strict disclosures rules could slow down local AI developers more than the global Big Tech.

And finally, we want to know how do you think about who should hold the power to verify?

So how can independent auditors access non-public data without concentrating power?

So it's open to any -- open the floor, so feel free -- please feel free to raise your hands and share your perspectives with us.

Anyone online or anyone?

>> Hi, I'm from Nepal. I come from a background in the technical [?] with experience in AI.

First of all, really good research from you guys as well. I want to bring up some points for Number 2, sovereign AI view.

So I think as your research suggested as well, companies are willing to take on the lawsuits and break a lot of rules if it guarantees that they're making a decent amount of money.

So I feel like imposing laws or imposing policies is only going to slow the -- like [?] developers who are trying to build in the AI space.

And for example, you mentioned EU quite a bit. EU is probably the weakest contender in this area, I'm pretty sure we all know. And I think people understand that GDPR is one of the reasons it is -- it is so.

Mistral is a great company but they are the weakest. They fail in comparison to the U.S. especially. Whilst there's a good mix of data protection and privacy that can be adopted from EU, I think if we go a little bit heavy on it, that will restrict local and Indy labs more than the frontier ones, because they're always, always willing to break the rules and take on the lawsuits after.

So I feel like there needs to be a well-maintained balance in that regard as well.

So thank you.

>> Hi, I'm from APrIGF team. So I expect the thought of [?]. Basically I'd like to say that big, you know, global tech companies look for just the profit they're getting out of it and not really thinking of the local communities or Indigenous communities which are underserved or underrepresented.

I think when we're talking about the disclosure rules, we must refrain from discussing the very viewpoint that we stick around the factor that global techs make in the tables without thinking about the underserved and then vulnerable people.

So they must take into consideration of that.

Thank you.

>> SODAM KUM: Thank you so much. I think we have -- we don't have enough time, so just one more person to say. Maybe here. Sorry. Is that okay?

>> Thank you so much for the opportunity and for the presentation. I think it's really nice.

Like for me, I'm actually curious because like most of the lows in the AI governance that we have had so far, I think it's mostly self-low. Which is in the form of guidelines.

Not only in Southeast Asia, I believe in Japan because I did research for Japan. Maybe Korea too, but you know better.

So my question like the AI sovereign that you mentioned before, when there is existing soft-low AI governance regulations in the countries, how are you going to create steps on how to create like control?

And your last question I think is interesting because I've been thinking government's supposed to be the first party to, you know, the first stakeholder to handle this.

But seeing the growing of AI adoption is far, it needs to be combinations of private sector, academia, Civil Society, everyone, they need to get together and discuss on these issues. All right.

So that is all.

>> SODAM KUM: So thanks for sharing your views and due to the time limit, we all have lots of limitations that we are facing right now, and I think it's our role to -- okay, it's our role to proceed what we have been living for the future.

So thank you.

>> I don't have question, I just have to add a little [?] who should hold the power to AI? As an end user, I think we should have tools to verify the AI content on our end. Maybe some tool or maybe some, you know, anything that we can -- that can help to us verify the content and everything.

>> SODAM KUM: Okay, guys. Thank you so much. Here is our mail address. So if you're interested or have questions about to reach us, please share it with us.

So yeah, it's like eg2ig2026ip@gmail.com. So like five, four, three, two, one, everyone get it? Okay. Thank you. Thank you so much.

>> Secretariat: Thank you, everyone. This is the secretariat speaking. Up next in this room, immediately in five minutes, is the session on community radio in the digital era. Strengthening Internet resilience and trusted communication in South Asia.

And in the next room, the next session in the next room will be who gets a say, a regional dialogue on digital public infrastructure, governance in Asia-Pacific.

That's next door. Next session.

And also don't forget always to add some of the highlights of this session into the Synthesis Document. You can go to comment.APrIGF.Asia to contribute your input. Do join us later at the town hall session at Lang Tang Hall which is the room next door for the town hall session for the Synthesis Document.

So we will break for about five minutes and reconvene.

(Session ended at 13:45 p.m. NPT)

This text, document, or file is based on live transcription. Communication Access Realtime Translation (CART), captioning, and/or live transcription are provided in order to facilitate communication accessibility and may not be a totally verbatim record of the proceedings. This text, document, or file is not to be distributed or used in any way that may violate copyright law.