

Tab 1

NISO Plus 2026 - Pre-Conference: Keeping the Robots in Line: Provisioning Access to Content for AI Systems - Collaborative Notes

[Pre-Conference - Keeping the Robots in Line: Provisioning Access to Content for AI Systems](#) - Monday, February 16, 2026, 8:30am - 1:00pm

- The primary consumers are machines, not humans
- We have only begun to build a world in which the robots will function well, as compared to the 30 years spent building a world for people to navigate digital information

Content Sourcing/Attribution: The Underlying Technology.... (Marjorie Hlava)

- Intelligent chunking at scale: Overlapping chunks, adding tags/more metadata to the top of a chunk, in-line tagging. Adding content from the “outside” to preserve the meaning results in domain-aware vectors. [there is a standard on this tagging - can someone add it? Marjorie is convener]
- Ideal infrastructure for publishers: Publisher Content to Chunk Pre-Processing to Vector Database to RAG to MCP (which enforces access control, licensing, etc.)

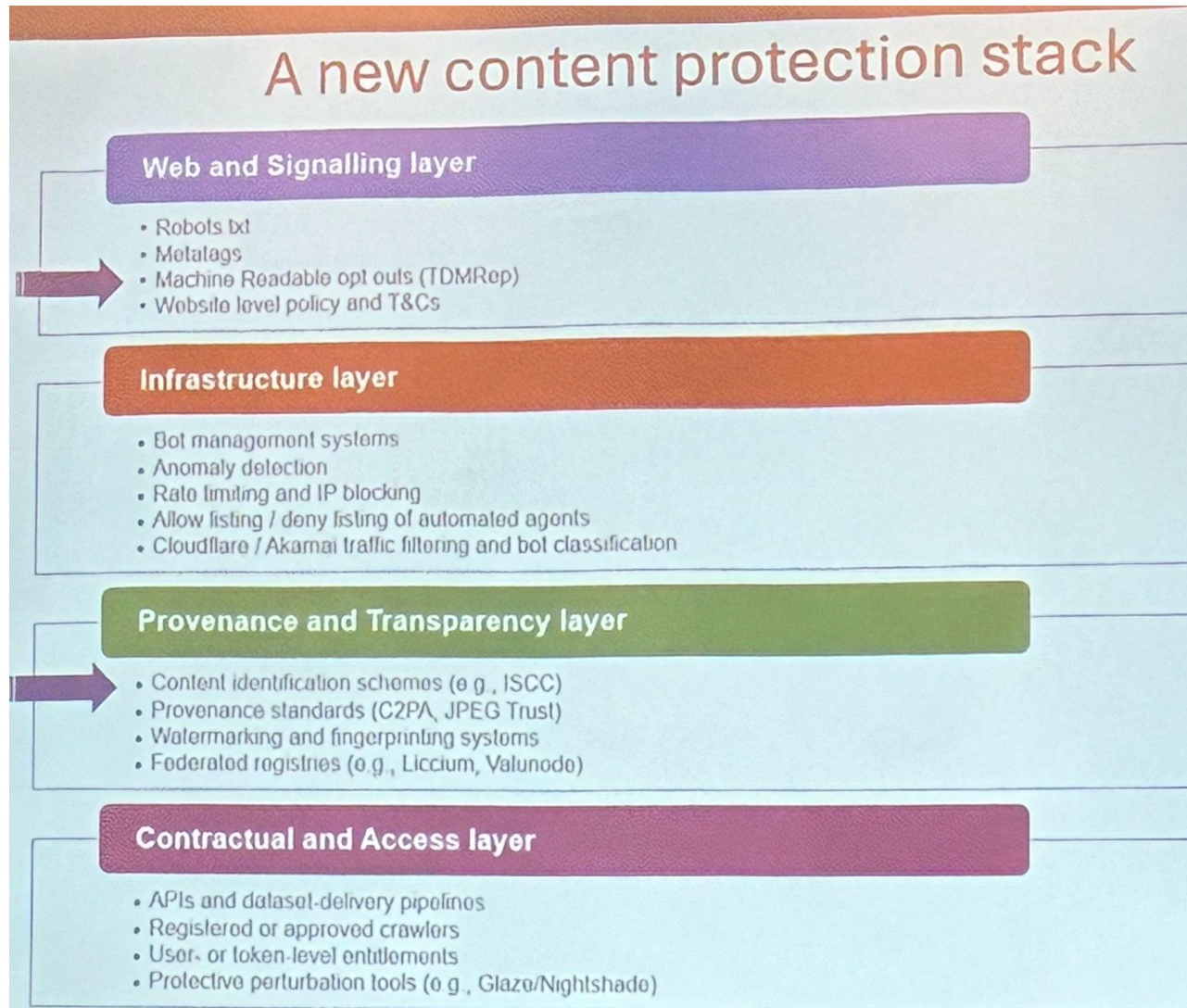
Business models, governance, and GenAI interoperability (Jessica Miles)

- Sustainable business and governance models as examples: Crossref, ORCID, Counter
- New models: Fees for training data, revenue sharing, usage and performance payments, institutional license integrations. A lot of deals are subject to NDAs, so there is little public information available. These deals are diminishing. Reduced need for training data. Unclear if revenue-sharing model results in sustainable revenue streams.

Publisher Perspective: Rights Recognition (Matteo Ripamonti)

- Natural Defences (pre-AI)
 - Robots.txt
 - IP-based
 - Licenses
- Changes
 - Scale

- Purpose
- Autonomy
- Incentives
- New protection stack
 - Web & Signalling
 - Infrastructure layer
 - Provenance and Transparency layer
 - Contractual and Access layer



<https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240202/>

- Article 3: Mandatory exception for scientific research
- Article 4: General exception for commercial TDM

Core parameters

tdm-reservation=1 (for TDM rights reserved)

Tdm-policy

- Implement by putting these tags into the meta name of the web pages.
- Redirects user to policy page which includes an email to request permission. Allows a dialogue with those who want to scrape the site content.
 - Sage is doing the same thing, but feels that systems are not going to follow up - has not seen a difference in usage
 - Elsevier has received about 100, 120 requests from the page
 - Is there even a “read” checkoff?
 - EU regulation requires binary, machine-readable instruction. Not sufficient to be in terms & conditions
-

Value

- > Can't monitor something you cannot measure
- > Search optimization - want to be in top three in Google Gemini vs. top of Google

From Access-Level Signalling to Post-Access Accountability ELSEVIER

TDMRep — Strengths

- Clear, machine-readable rights signaling at the point of access
- Legally grounded (DSM Directive, Article 4)
- Simple, binary, interoperable
- Works well for good-faith actors & website
- Creates a clear contact and licensing path

TDMRep — Structural Limits

- Hard-binding: signal lives where content is served
- Does not travel reliably once content is downloaded or redistributed
- Limited coverage of historical backfile
- Not designed for detection, provenance, or post-hoc accountability
- Cannot resolve “we didn’t know” claims after the fact

• We need content-level identification that works after access

• Signals must be resolvable independently of platform or URL

ISCC - ISO standard to identify documents

Does this take into account identifiers from LLMs, eg, <https://deepmind.google/models/synthid/> ?
No, but worth investigating?

Fundamental difference between ISSC and DOI

Use-case: Photography Association in Germany

Is there a community driver advocating for adoption?

Similar to blockchain. We have the technology, do we have a use case? You do not need a blockchain to do the ISSC identification and registry

<https://registry.commonsdatab.org/explorer>

- Wikimedia took a different approach - the idea that if there is an ICSS, then the content is free to use

Keeping the Robots in Line: Product Development Perspective (Tony Alves)

- How do we design systems that are future resilient?

AI Agents are “not polite”?

<https://www.researchinformation.info/analysis-opinion/after-the-pdf-a-new-unit-of-knowledge-for-the-ai-era/>

- AI agents are already here and they are not polite
- They are arriving at scale
- PDFs and articles are the wrong unit for AI
 - They need claims, figures, methods, and assertions, without context
- Distinguish: training versus grounding/RAG versus inference
- Encode provenance - context and provenance matter more
- “Rights metadata becomes core infrastructure, not legal boilerplate”

Entitlement Aware Access

No user > no entitlement

Controlled API Gateways

APSI become the primary access route

Verifiable Agents -

Intermediated RAG Access - minimize delivery of raw content

MCP - Model Context Protocol

Controlled Intermediary Access Layer

Agent-to-Agent Communication

Secure Training Environments

We are talking about an infrastructure shift, so that AI can be a governed participant in the ecosystem, not only a consumer

“Request, verify, deliver” instead of “scrape first, ask questions later”

“If AI agents are going to become first-class consumers of scholarly content, what do we need platforms to guarantee: access control.... (and more!)”

Tim Lloyd

Bots do not behave the same way as humans. So, you can tell the difference.

Many bots for a vendor that are also using CloudFlare (with low settings)

Liblynx conducted bot tracking “Bot Traffic Analysis” and created bot identification from the activity

IDEAS:

Create a taxonomy of bots and bot behavior

Notes from TMC, added 3-March-26

- Move to machine-first reading means quality metadata is more important than ever
- New services need decent governance and proper sustainable business models. Fees for training data are drying up (training is mostly complete), but revenue sharing based on usage and performance payments, and/or institutional license integrations, are still out there. A lot of deals are subject to NDAs. It's unclear if any of these models will be sustainable in the long term.
- Bots don't behave: 80% of bots don't declare themselves properly on platform visits. We can extend COUNTER's bar analogy: a bar allows everyone in until they behave poorly. A club only allows you in if you have the right ticket. Do publishers need to start authorising first (giving out tickets) rather than rejecting when behaviour is poor?
- Rights recognition is changing. AI tools may ignore robots.txt (not surprising). They operate on a different scale and with a level of autonomy that trad bots didn't have. **We can't monitor what we can't measure.** We have traditionally relied on web signalling (e.g. website policy) to protect us. Adding infrastructure (e.g. Cloudflare) is newer, but we need to add a new contractual / access layer such as data delivery pipelines to better work with AI.
- “We are talking about an infrastructure shift, so that AI can be a governed participant in the ecosystem, not only a consumer.”
- Bots are not human, so we can distinguish between them. Trad bots and AI agents have different purposes: scrape versus summarise. Liblynx has a bot traffic analysis tool that allowed them to create a taxonomy of bot behaviour