

Presta, Marco M [NC]

From: Presta, Marco M [NC]
Sent: October 17, 2025 1:44 PM
To: Addardouri, Mehdi MEMA [NC]; Gupta, Atul [NC]
Subject: Emerging risk: Authorized Access to Appeals & Decisions for EI Jurisprudence Project

Follow Up Flag: Follow up
Due By: October 30, 2025 10:00 AM
Flag Status: Flagged

Hi Mehdi, Atul

Perhaps I am precipitating it but I raised the risk based on our chat 2-3 days ago and suggestion to have the client seeking those official agreements...

Regards
Marco

From: Presta, Marco M [NC]
Sent: October 17, 2025 12:28 PM
To: Lavergne, Eric EL [NC] <eric.lavergne@hrsdc-rhdcc.gc.ca>; Whitley, Todd TW [NC] <todd.whitley@hrsdc-rhdcc.gc.ca>; Blake, Niasha N [NC] <niasha.blake@hrsdc-rhdcc.gc.ca>; O'Rourke, Shannon SM [NC] <shannon.orourke@servicecanada.gc.ca>
Subject: fyi: Emerging risk: Authorized Access to Appeals & Decisions for EI Jurisprudence Project

Hello All

Quick background on the data needed by Jurisprudence

- **Primary corpus (now): SST appeal decisions** published on the Social Security Tribunal website (public). These power EI case-law lookups in Jurisprudence.
- **Planned additions: Federal Court decisions** (new pipeline + index) to reflect hierarchy/precedence across tribunals and courts.
- **Terms-sensitive source (under review): CanLII** content flagged for potential **licensing/terms-of-use restrictions**; governance required before inclusion.
- **Future enrichment (nice-to-have):** Link to an **annotated EI Act** and show **hierarchy of decisions** (e.g., which decision supersedes), improving citation quality.
- **Data engineering posture:** Build **incremental ingestion** for all sources to handle updates/takedowns and maintain **data history** for audit.

1) Risks with website scraping

- **Legality / Terms-of-use uncertainty:** Public ≠ freely reusable. Site terms (and robots rules) may restrict automated collection or reuse; similar licensing caveats are already flagged for other sources like CanLII.

- **Fragility & reliability:** Scrapers break when the site structure, pagination, or URLs change; CAPTCHAs/throttling can intermittently block access; availability outages stall ingestion. (PoC relied on decisions from the SST website, which reinforces this exposure.)
- **Incomplete metadata & auditability:** HTML pages often lack stable identifiers, revision history, or takedown flags; that weakens provenance, deduping, and traceable citations.
- **Load & ethics:** High-frequency crawling risks overloading SST infrastructure and can contravene good-citizen crawling norms.
- **Security / compliance:** Uncontrolled parsing increases attack surface (e.g., HTML payload oddities); harder to prove end-to-end compliance in audits.

2) Risks when updating previously scraped data

- **Data drift & regressions:** If SST revises, republishes, or withdraws a decision, scrapers may miss it or create duplicates; the search index can drift from the source of truth. (Our scope explicitly calls for **incremental pipelines** to combat drift & costs—good, but scraping remains brittle.)
- **Change detection gaps:** Without authoritative feeds (IDs, Last-Modified, changelogs), we rely on brittle heuristics (hashing, diffs) that can miss subtle edits (e.g., anonymization updates).
- **Timeliness vs. load trade-off:** Polling often to stay fresh increases site load/risk of blocking; polling less risks stale content.
- **Versioning & traceability:** Hard to maintain an auditable “what was shown to users on date X” without canonical IDs and official version metadata.
- **Multilingual & structure shifts:** Layout or taxonomy changes (topics/issues) can silently degrade extraction quality and ranking.

I raised it to the PM in the email below and suggested the client reaching to content owners for official and reliable data extraction mechanisms.

[@O'Rourke, Shannon SM \[NC\]](#), [@Blake, Niasha N \[NC\]](#) please chime in.

Thanks

Marco

From: Presta, Marco M [NC]

Sent: October 16, 2025 11:06 AM

To: Addardouri, Mehdi MEMA [NC] <mehdi.addardouri@hrsdc-rhdc.gc.ca>; Gupta, Atul [NC] <atul.gupta@hrsdc-rhdc.gc.ca>; O'Rourke, Shannon SM [NC] <shannon.orourke@servicecanada.gc.ca>; Groulx, Kris [NC] <kris.groulx@hrsdc-rhdc.gc.ca>

Subject: draft: Inquiry on Authorized Access to SST Appeals Decisions for EI Jurisprudence Project

Hi Mehdi, team

Would you ask the client to do the inquire the data providers, please?

draft

Hello Social Security Tribunal Team,

I'm writing from Employment and Social Development Canada (ESDC).

We are developing the **Employment Insurance (EI) Jurisprudence** solution—an internal, bilingual research tool to help EI adjudicators quickly find and cite SST appeal decisions.

During a proof-of-concept phase led by a contractor, decisions were collected via a web-scraping approach from public SST pages.

As we move to a production-grade solution operated by ESDC, we want to **discontinue scraping** and instead obtain an **authorized, stable, and terms-compliant** method of accessing SST appeal decisions.

Specifically, we'd appreciate your guidance on:

1. Preferred Access Mechanism

- Is there an **API**, bulk **data export**, **RSS/sitemap**, or other supported feed for decisions and metadata?
- If not available today, are there **approved guidelines** for automated retrieval that respect your infrastructure and usage policies?

2. Terms, Licensing, and Attribution

- Any **Terms of Use** or licensing restrictions governing reuse by Government of Canada.
- Required **attribution/citation** format for SST decisions.
- Any **robots.txt** or rate-limit rules we should enforce, if automated retrieval remains the interim method.

3. Scope and Coverage

- Data elements available (e.g., docket/decision number, appeal division, member, date issued, parties anonymization, topics/issues, outcomes, language, PDF/HTML links).
- Coverage for **historical** decisions and ongoing **update frequency** (e.g., daily/weekly).
- Any **retentions/withdrawals** we must reflect (takedown/change notifications).

4. Operational Considerations

- Contact point for a **data-sharing agreement** (if required).
- Expected **service levels**, planned website changes, or **deprecation** schedules that could affect integration.
- Any **costs** associated with access.

We're committed to complying with SST policies, minimizing load on your website, and ensuring proper attribution and data stewardship. If helpful, we can share a short architectural note describing how the data will be stored and used internally at ESDC.

Could we schedule a brief call (30 minutes) to discuss options? I'm happy to accommodate your availability this week or next.

Thank you for your time and support.

Marco Presta

Artificial Intelligence Centre of Enablement (IITB/AICoE)
Employment and Social Development Canada / Government of Canada
marco.presta@hrsdc-rhdc.gc.ca

Centre d'Habilitation de l'Intelligence Artificielle (DGIIT/CenHIA)

