Statement on Protecting News Content in Generative AI

June 4, 2025

Japan Newspaper Publishers & Editors Association

The Japan Newspaper Publishers & Editors Association has consistently advocated that generative AI services obtain consent from copyright holders when utilizing or training on news content. While there are insufficient rules for content protection, setting up technical measures like robots.txt is an extremely effective aid. This allows rights holders to easily and clearly express their intent to refuse unauthorized training or utilization of news content, while imposing minimal burdens on those engaging in data training or utilization. Given this, it is only natural that businesses developing generative AI, providing services, or collecting data should respect rights holders' intentions. If they intend to utilize or train on news content, getting permission from rights holders is the proper course of action. At the same time, we urge the government and Diet to expedite the establishment of systems for the appropriate protection of content.

AI Businesses Should Comply with robots.txt

Major news sites operated by our association's member companies have set up robots.txt to indicate their intent to protect content. Most major domestic portal sites that feature articles provided by newspapers and news agencies have similar settings. Article 30-4 of the Copyright Act, which governs the use of copyrighted works for AI training, contains no explicit opt-out provision allowing rights holders to refuse use, nor does it contain explicit arrangements regarding technical measures. However, the General Understanding on AI and Copyright in Japan presented by the Agency for Cultural Affairs in March 2024 clarifies that when a database—containing data from websites on the internet and organized in a form usable for information analysis—is sold, or its sale is reasonably expected, circumventing technical measures to collect content for AI training purposes could constitute copyright infringement under the proviso of Article 30-4, which prohibits acts that "unreasonably prejudice the interests of the copyright holder."

Recently, services using Retrieval-Augmented Generation (RAG) technology—where AI generates responses associated with web searches—have rapidly expanded. The Agency for Cultural Affairs' General Understanding states that if answers generated by such services utilize the original

work beyond the scope of minor use defined in Article 47-5 of the Copyright Act, the copyright holder's permission is required. The enforcement regulations concerning Paragraph 1 of the same article stipulate that the minor use provision does not apply if collection of the original work is prohibited by means such as robots.txt. Therefore, operating such services by circumventing technical measures carries an elevated risk of constituting copyright infringement.

Several businesses ignore robots.txt while collecting data. We confirmed articles from member companies of the Newspaper Publishers Association appearing as references even though they had set up robots.txt. This is an unacceptable situation. When rights holders demonstrate their intent to protect content by setting up robots.txt, AI businesses must comply with it for both training and utilization.

Ensuring the Effectiveness of robots.txt

Another problem is that some AI businesses conceal the information needed to implement technical measures, making it impossible to prevent AI from using or training on news content.

Using robots.txt requires the name of the crawler that is collecting content (user agent information), yet many businesses collect data without disclosing the crawler name. If data collection by businesses who hide their crawler names is left unchecked, the effectiveness of the content protection methods outlined in the Agency for Cultural Affairs' aforementioned General Understanding cannot be guaranteed, making it difficult for rights holders to take countermeasures. Article 47-5 of the Copyright Act recognizes opt-out via robots.txt and similar methods under its enforcement regulations, which could curb unauthorized use of news content by RAGs. However, similar problems arise when user agents are invisible. It has also been pointed out that AI businesses may purchase data collected by other entities that do not disclose their user agents, rather than collecting content directly themselves.

It is imperative to mandate the disclosure of user agents—essential for configuring robots.txt—not only for AI businesses but for all data collection entities. Furthermore, a system enabling rights holders to readily access this disclosed information must be established urgently.

Additionally, the lack of distinction between crawlers for generative AI and those for search services prevents rights holders of news content from expressing their intentions appropriately. It is almost impossible for rights holders to allow crawlers for search services to guide users to news content while blocking crawlers for generative AI. To ensure adequate protection of news content, standardizing protocols that allow rights holders to appropriately express their intentions—such as establishing technical methods to block only crawlers for generative AI—is essential.

Calling for Comprehensive Measures to Protect News Content

The government has indicated in its guidelines for the new AI law that it will consider providing information to ensure dataset transparency. Guidelines for businesses issued by the Ministry of Internal Affairs and Communications and the Ministry of Economy, Trade and Industry list the appropriate collection of training data as a key requirement for AI businesses. While the Agency for Cultural Affairs' General Understanding has advanced the clarification of concepts regarding copyright and AI to a certain extent, it must be said that the protection of news content remains extremely precarious.

Generative AI technology advances daily, and the problem of "zero-click search"—where many users are satisfied with AI-generated answers and do not visit the source website—is getting worse and worse. New services that free-ride on news organizations' content continue to emerge, such as "deep research" tools that analyze large volumes of articles to generate answers and platforms that let users republish article collages as websites.

If this problem continues, the cycle of content reproduction could be disrupted, potentially forcing news organizations to scale back their reporting operations. Generative AI will never replace news organizations in conducting interviews and reporting. A decline in the role of news organizations, which are vital contributors to news dissemination, could undermine the public's right to know. This is a critically important issue concerning the very foundation of democracy. We demand comprehensive responses that extend beyond current frameworks such as copyright and competition law.