



**BEFORE THE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY**

**Guidance and Templates for Public-Facing
AI Documentation: An AI Standards 'Zero
Draft'**

NIST AI 300-1 ipd

COMMENTS OF THE COPYRIGHT ALLIANCE

The Copyright Alliance appreciates the opportunity to submit the following comments in response to the [call for input](#) by the National Institute of Standards and Technology (NIST), regarding its initial public draft of "Guidance and Templates for Public-Facing AI Documentation: An AI Standards 'Zero Draft,'" aimed at piloting a process to broaden participation in and accelerate the creation of standards to meet the AI community's needs and unleash AI innovation.

The Copyright Alliance is a non-profit, non-partisan public interest and educational organization representing the copyright interests of over 2 million individual creators and over 15,000 organizations in the United States, across the spectrum of copyright disciplines. The Copyright Alliance is dedicated to advocating policies that promote and preserve the value of copyright, and to protecting the rights of creators and innovators. The individual creators and organizations

that we represent rely on copyright law to protect their creativity, hard work, and investments in the creation and distribution of copyrighted works for the public to enjoy.

The Copyright Alliance and its members are on record as supporting generative AI, so long as it's developed and used in a manner that is responsible, respectful, and ethical.¹ Adequate and appropriate transparency and record keeping are essential elements of each of those benchmarks. From a copyright perspective, transparency of ingestion of copyrighted works by businesses that offer generative AI systems to the public is critical because it will help ensure that the rights of copyright owners are respected, as it allows copyright owners to identify if and when their works have been used for training.

The Copyright Alliance and its members support the establishment of a process to broaden participation in and accelerate the creation of standards to meet the AI community's needs and unleash AI innovation, and we believe the Zero Draft is a valuable step toward standardized AI documentation. Specifically, the Copyright Alliance supports transparency (through both voluntary standards and legal obligations) surrounding the copyrighted works used in generative AI training and the documentation of AI datasets and models for transparency between AI stakeholders, subject to certain limitations discuss in detail below.²

Why Transparency is Important

At the time of this writing, there have been over 140 copyright infringement suits brought by copyright owners against generative AI companies. Complicating these lawsuits is the fact that AI companies are increasingly secretive and less transparent about the copyrighted works ingested for training their AI systems. AI companies are the only parties who know what is in their training data; copyright owners cannot access this information any other way. Thus, unless

¹ Artificial Intelligence, Copyright Alliance, <https://copyrightalliance.org/policy/positions/artificial-intelligence/> (last visited Aug. 26, 2026).

² These comments address transparency related to training data, not output labeling.

copyright owners can demonstrate copying through evidence of verbatim or near verbatim outputs, they must rely on litigation discovery to identify whether their works were used for training. However, discovery is inefficient, costly, slow, and may not even result in disclosure of the relevant information. Requiring copyright owners to shoulder this burden will result in AI companies' continued exploitation of copyright owners and creators, especially independent creators.

Importantly, adequate and appropriate transparency and record-keeping benefit both copyright owners and AI developers in resolving questions regarding infringement, fair use, and compliance with licensing terms. Additionally, transparency has many other benefits unrelated to copyright such as promoting safe, ethical, and unbiased AI systems that support scientific reproducibility and validation.³ Consequently, there can be little doubt that transparency by businesses that offer generative AI systems to the public is a crucial component of any AI policy that will ultimately result in safer and more ethical technologies.

Scope and Applicability of Transparency Requirements

Transparency standards should be developed to ensure AI companies maintain records of what copyrighted works are being ingested, how they were obtained, how those works are being used, and whether the works reside or are stored somewhere within the system or elsewhere. Importantly, disclosure of this information should be limited and only apply when each of the following criteria is met:

- An entity or individual that develops generative AI systems by ingesting material that they do not own or have not licensed for training purposes.
- An entity or individual that *offers generative AI systems to the public or disseminates systems to third parties who subsequently offer the systems to the public.*

³ These comments will not detail the benefits unrelated to copyright, as they are outside of the Copyright Alliance's mission and expertise.

- When such obligations would not be contrary to or inconsistent with obligations under other laws, contracts, or collective bargaining agreements.

The licensing/owned works exclusion above serves several purposes. First, there would be no risk of infringement if the AI developer retained all rights that could be implicated by ingesting and copying works as part of training a model. Second, the AI developer should not have any obligations where a work is licensed because if the licensor has concerns with how their work is to be used and protected, those can be addressed in the terms and conditions during the license negotiations. Third, presumably the licensor-copyright owner is compensated for and has authorized the use of their works so there would be no “free-riding” concerns. Fourth, there is no reason for the transparency obligations to apply because the licensor-copyright owner is presumably aware that the AI company is ingesting their copyrighted works. And finally, and perhaps most importantly, for those AI companies that oppose transparency, it gives them a simple solution to address their transparency concerns—if they license the work, the transparency obligations would not apply to those works.

Section 1 of the Zero Draft states that the document is “applicable to public-facing documentation of datasets and models whether or not the datasets and models themselves are publicly available.” As noted above, the Copyright Alliance believes that disclosure requirements should be limited to entities or individuals that offer generative AI systems to the public. Thus, any disclosure recommendations outlined by the Zero Draft for entities that only utilize generative AI models internally should remain voluntary.

As the Zero Draft explains in Section 3.4.4 on Legal Obligations, “[d]ocumentation processes can contribute, to varying degrees, to successful or unsuccessful fulfillment of obligations under relevant laws, regulations, or contracts (e.g., transparency or reporting requirements or obligations regarding data privacy or security).” This position aligns with the Copyright Alliance’s third point above that disclosure requirements or recommendations must consider and not be inconsistent with obligations under other laws, contracts, or collective bargaining agreements.

Finally, transparency standards should apply to materials used for training and not any trade secrets, details about weights or parameters used by the AI company in constructing a generative model, algorithms, or anything else that would reveal sensitive proprietary information. Some AI companies claim that requiring transparency about the material they use to train their AI systems will require them to divulge their trade secrets and doing so would extinguish any trade secret protection that exists in their training methods or AI models, including proprietary algorithms. But the copyright community is not asking for such information. What the copyright community wants is for AI companies to disclose the copyrighted works that they use to train their AI system, subject to the limitations listed above.

In effect, the copyright community is only asking for AI developers to disclose the copyrighted works used to train their system without a license, which typically means those that were scraped off the internet. As evidenced by the definitions above, there is no trade secret protection in the collection of these works because they are publicly available to anyone who has internet access. There also wouldn't be trade secret protection for the process of scraping the entire internet for material, as it is a generally known and ascertainable way to create a training dataset. Moreover, many AI companies are using the same sources (e.g., common crawl, Books3) for these materials and most AI developers see their competitive advantage in their algorithms, and do not view choice of training materials as a source of competitive advantage.

Considerations Specific to Certain Types of Copyrighted Works and Copyright Owners

A. Scientific Works

When considering transparency recommendations, the Zero Draft should emphasize the importance of the use and disclosure by AI developers of the version of record of copyrighted scientific works. Versions of record, which are typically provided directly by the rightsholder or

its agent, consist of verified, peer-reviewed versions of articles and other literary works.⁴ It is critical that AI training corpuses used for scientific research purposes consist of these versions of record and not preprints, drafts, or versions where changes have been made subsequent to posting.

Without guardrails that ensure the use of versions of record, the training of high-stakes research models (such as those used in pharmaceutical research) would be compromised by the inclusion of inaccurate or out-of-date content, as well as content sourced from predatory journals and/or “junk” science.⁵ When such content is ingested for AI training, the model risks generating output that is biased, inaccurate, and/or wholly hallucinated. Advancements in transforming scientific research through AI can only be fully realized if AI development proceeds on a foundation of respect for IP rights, fair compensation for copyright owners, and transparent, accountable use of training data. Ultimately, transparency standards for AI training are not obstacles to innovation but rather essential measures that will ensure sustainable, ethical, and legally sound AI development.

B. Music

Transparency surrounding musical works used for generative AI training is of particular importance due to the nuances regarding separate copyrighted works contained in a song or sound recording and the need to identify training material on a work-by-work basis. Unlike many other categories of copyrighted works, a single piece of recorded music typically embodies at least two distinct copyrighted works: the musical composition (the underlying song, including the music and lyrics, generally owned by the songwriter(s) and/or music publisher) and the sound recording (the fixed performance of that composition, generally owned by the recording artist and/or record label). These two copyrights are protected separately under the Copyright

⁴ Lisa Janicke Hinchliffe, *The Version of Record as a Central Organizing Concept in Scholarly Publishing*, 42 *Info. Services & Use* 63 (2022) (explaining that the version of record serves as the fixed, publisher-designated reference point around which all other article versions, corrections, and citations are organized).

⁵ Roy Kaufman, *Swimming in the AI Data Lake: Why Disclosure and Versions of Record Are More Important than Ever*, *The Scholarly Kitchen* (May 15, 2023), <https://scholarlykitchen.sspnet.org/2023/05/15/swimming-in-the-ai-data-lake-why-disclosure-and-versions-of-record-are-more-important-than-ever/>.

Act, and each may be infringed independently. A single sound recording may implicate additional layers of rights and interests as well, including the contributions of featured artists, session musicians, and/or background vocalists, each of whom may hold a distinct, independently enforceable interest in the same recording.

Because the composition and sound recording embodied in a single track are often owned and administered by entirely different parties—including composers, publishers, performers, and labels—and because a single recording may combine elements drawn from multiple other works, a general disclosure that "music" was used for training provides little practical value to the many copyright owners. Work-by-work transparency is therefore essential to allow each affected rights holder to determine whether and how their specific copyright interest was used. For example, copyright owners should be able to determine:

- Whether the musical composition, the sound recording, or both were ingested for training;
- The specific title, songwriter(s) and/or performing artist(s), and standard industry identifiers;
- Whether featured artists, session musicians, or other contributors are separately identified in connection with the ingested recording.

If work-by-work level information cannot be publicly disclosed (due to one or many of the limitations listed above), the Zero Draft should require a mechanism for obtaining such information for legitimate purposes, such as by rights owners or creators wanting to know if or how their works were utilized.

C. Individual Creators Concerns

While the implementation of appropriate transparency standards is important for all copyright owners, it is of particular significance to individual creators and small businesses that do not have the resources to prove through litigation that their works were used for training. Indeed, transparency is not simply a first step towards filing a lawsuit, but rather it also facilitates

licensing and other business solutions that ensure creators can control or be compensated for use of their works. When individual creators and copyright owners are in the dark about when or whether their works have been used for training, they have very little recourse or leverage to enforce their rights and/or work towards licensing systems.

Conclusion

It is vital that transparency standards be implemented to ensure that AI development is responsible, respectful, and ethical, and that it serves the interests of copyright owners and AI developers alike. The Copyright Alliance urges NIST to ensure the Zero Draft supports the establishment of publicly accessible, searchable records documenting what copyrighted materials are used to train generative AI systems and how those materials are used. Such transparency standards should (1) apply only to developers that offer generative AI systems to the public, (2) not extend to works that are owned or licensed for training, and (3) not conflict with obligations under other laws, contracts, or collective bargaining agreements.

The Zero Draft should also recognize the particular importance of the version of record for AI systems trained on scientific and scholarly works, so that high-stakes applications are not compromised by inaccurate, outdated, or predatory-journal content. Additionally, the draft should recognize the complexities surrounding copyright ownership in musical works and ensure that transparency standards apply to all relevant rights holders. Lastly, the Zero Draft should recognize that transparency measures are especially critical for individual creators and small businesses, which lack the resources to establish through litigation that their works were used without authorization.

The Copyright Alliance appreciates NIST's efforts to accelerate the development of AI standards through this pilot process and welcomes the opportunity to continue engaging as this work progresses.

Respectfully Submitted,

A handwritten signature in black ink, appearing to read 'KK', with a long horizontal flourish extending to the right.

Keith Kupferschmid
CEO
Copyright Alliance
1331 F Street, NW, Suite 950
Washington, D.C., 20004

September 16, 2026