STM Advancing trusted research

January 14, 2026

Hal Finkel
Office of Science
U.S. Department of Energy
1000 Independence Avenue SW
Washington, DC 20585-0121

Submitted via email to AICommunityInput@science.doe.gov

To the DOE Office of Science AI Team:

STM, the International Association of Scientific, Technical and Medical Publishers, welcomes the opportunity to respond to the Department of Energy's Request for Information (RFI) on Partnerships for Transformational Artificial Intelligence Models (DE-ASCR-26-0001). STM and its members are both intensive users and developers of advanced AI tools, and they are also primary providers and stewards of high-quality scientific and technical content that underpins trustworthy AI systems and the broader U.S. research enterprise.

STM's interest and overarching perspective

STM represents a global community of scholarly and professional publishers who invest in editorial processes, peer review, integrity checks, metadata standards, and long-term curation to ensure that scientific and technical information is accurate, traceable, and reusable. STM strongly supports the Administration's focus on building American AI infrastructure and scientific leadership, as reflected in Winning the Race: America's AI Action Plan, and believes that regulatory and partnership frameworks should remove unnecessary friction while preserving the policy foundations—copyright, licensing, transparency, and integrity safeguards—that have long enabled U.S. leadership in science and innovation.

For the DOE AI consortium to achieve its aims—self-improving AI models that accelerate discovery, advance energy innovation, and strengthen national security—it must be grounded in lawful access to high-quality, validated content and data, combined with strong protections for intellectual property and research integrity. With respect to protection and promotion of IP, STM supports and endorses the submission of the Copyright Alliance, which goes into significant depth on the importance of respect for IP and copyright in DOE policy.

Existing copyright and licensing regimes provide the necessary incentives and guardrails to support accurate, reliable research findings and trustworthy AI. As noted in our recent submission to OSTP on "Accelerating the American Scientific Enterprise," AI-enabled science needs to build on validated research findings and have transparency of training inputs to prevent inaccurate information that can harm public safety and understanding and create inefficiencies. Here, we note that "training data" as it is used in the RFI not only encompasses data associated with copyrighted works, but the works themselves. Because there are significant legal and rights differences between copyrighted content and data, STM prefers the term "training input." STM

recommends whatever terms are used are accompanied by clarifications that highlight the important differences between the two different types of content to ensure that the rights in content are respected in both policy and practice. This theme is further developed in our responses below and with recommendations in the appendix. More details are also articulated in our response to the OSTP RFI on the AI R&D Strategic Plan.

Importance of accuracy and integrity for science

Across all aspects of the consortium—data curation, model development, evaluation, and deployment—accuracy and integrity must be treated as central design goals, not optional features. High-quality peer review, editorial oversight, and post-publication correction mechanisms are indispensable for maintaining a reliable scientific record; AI models that are trained or evaluated without regard to this curated record risk amplifying errors, biases, and fabricated content at scale.

STM therefore urges DOE to: (1) prioritize licensed, peer-reviewed, and corrected literature and high-quality datasets as primary inputs for scientific and engineering AI models; (2) prioritize accuracy and integrity in development and deployment of the consortium, in coordination with private-sector experts, where appropriate; and (3) promote transparency and verifiability mechanisms—such as citations, provenance metadata, and audit trails—that allow researchers and the public to understand how AI systems arrive at their outputs.

Critically, scientific publications are complex copyrighted works that embody creative expression, selection, and arrangement, and they are subject to well-established intellectual property frameworks and licensing markets; they cannot be treated as generic "data" without undermining those frameworks and the quality and integrity of the works that contribute to American scientific advancement. While efforts to accelerate AI development by ignoring copyright and IP may seem attractive in the short term, policies that do so will erode U.S. competitiveness, and associated innovation and jobs. Weakening or bypassing existing IP and transparency frameworks, whether through over-broad definitions of "data" or licensing overrides, would, contrary to the Administration's goals, erode incentives for human creativity and discovery and allow synthetic, unvetted material to overwhelm the high-quality content that has made U.S. science and publishing a global leader.

Mobilizing data and labs while protecting rights (RFI Q1)

DOE's plan to curate scientific data across the National Laboratory complex offers a unique opportunity to combine high-value public sector data with the peer-reviewed literature and related outputs that publishers steward, creating AI-ready corpora that are both powerful and trustworthy. STM encourages DOE to structure curation and preprocessing work so that it is tightly coupled to FAIR (Findable, Accessible, Interoperable, Reusable) and "AI-ready" practices, including persistent identifiers, rich metadata, explicit license information, and documented provenance for all datasets and textual inputs. These features are enabled and supported by publishers' investments, including work on reliable and useful data sharing, and accurate licensing.

To mobilize the National Laboratories effectively, DOE should: (1) establish common data and content governance frameworks across labs that recognize existing copyright and licensing

arrangements for scientific publications and datasets; (2) promote lawful, licensed access to publisher content for training and evaluating AI models; and (3) work with publishers and existing standards to embed research integrity and quality criteria into data-curation workflows to prevent low-quality, synthetic, or unvetted material from polluting high-value training sets.

STM and its members support protection for sensitive or proprietary data and note that privacy and security protections should operate alongside copyright and licensing compliance, not as substitutes for it.

Structuring the consortium and combining models (RFI Q2)

STM recommends that DOE's public-private consortium be structured to recognize distinct but complementary roles: DOE and its laboratories as providers of mission-relevant raw data, facilities, and domain expertise; AI developers as builders of general-purpose and scientific models; and publishers as suppliers of curated, high-quality and verified peer-reviewed works and integrity infrastructures. Engagement with scholarly publishers and related infrastructure providers (for example, cross-publisher research-integrity initiatives) will help the consortium prioritize accuracy, traceability, and lawful use of content from the outset.

For general purpose AI models, STM recommends finetuning or adapting such models using licensed, high-quality scientific content—particularly final Versions of Record and corrected literature—to strengthen reasoning, hallucination-free accuracy, and support reliable domain specific inference. Access to publisher content should occur only under negotiated licenses with appropriate technical controls. Architectures built on authoritative, licensed corpora ensure outputs remain grounded in verifiable, up to date sources and enable clear pathways for citations, provenance, and links back to the underlying record of science and the entities that take responsibility for the integrity of the record. DOE might look to existing public-private consortium frameworks like the National Institute of Health's Generalist Repository Ecosystem Initiative (GREI) for models that enable access within a collaborative, IP-protecting, distributed framework.

Model evaluation should place particular weight on scientific accuracy, calibration, and reliability, rather than only on generic benchmarks. STM encourages DOE to incorporate evaluation methods that compare AI-generated outputs against peer-reviewed literature and established domain benchmarks, use human expert review in critical domains, and assess whether models appropriately highlight uncertainty and limitations in line with the Administration's Gold Standard Science guidance.

Providing models via cloud and research infrastructure (RFI Q3)

DOE's intention to provide AI models to the scientific community via cloud-based programs and infrastructure can significantly accelerate innovation if access models are designed to support transparency, reproducibility, and respect for rights. Cloud environments should enable researchers to see what data sources, licenses, and provenance are associated with a given model, and provide tools for surfacing citations and links to publications and datasets when model outputs rely on specific sources. Such features will encourage the use and reliance on tools that are trustworthy by researchers and the public.

STM encourages DOE to: (1) incorporate robust provenance and licensing metadata within any hosted models; (2) ensure that usage policies for DOE-provided models are consistent with the licensing terms under which training data and content were accessed; and (3) support mechanisms that provide licensed access and licensing requests directly within DOE-aligned cloud environments. Such approaches would support a dynamic licensing market for AI training and deployment, consistent with STM's broader AI policy work and the principles advanced by the Creative Rights in AI Coalition.

Cloud platforms are also well positioned to support integrity-enhancing services, such as cross-publisher checks for manipulated images or paper-mill content, which STM and its members are already developing through initiatives like the STM Integrity Hub. Integrating such services into DOE-supported research workflows would help ensure that AI-driven discovery is built on a reliable evidentiary base, reinforcing the accuracy and trustworthiness of both models and the scientific record.

Governance, FAIR/AI-ready repositories, and federated data (cross-cutting questions 1–5)

Effective governance for shared data, models, and infrastructure must address legal compliance, research integrity, and operational accountability in a unified way. STM recommends that DOE adopt governance models that: (a) clearly distinguish between public-domain, open-licensed, and restricted content; (b) require documentation of legal bases (including contracts and licenses) for any non-public material used in training; and (c) embed expectations for transparency, provenance, and integrity into consortium norms, not only project-specific agreements.

Preparing scientific data at scale for AI training requires not just technical standardization but also alignment with the publishing system that communicates and validates those works. STM urges DOE to coordinate with publishers, repositories, and identifier systems to promote consistent use of DOIs and other persistent identifiers, rich metadata (including funding and licensing information), and standardized links between publications, underlying data, software, and preprints, thereby creating AI-ready repositories that reflect both FAIR and "trustworthy AI" principles. STM's data sharing guidelines could be leveraged to support the utility of data for AI.

On the question of centralized versus federated or distributed repositories, STM sees a need for a balanced approach. Centralized catalogues or indices can provide unified discovery, metadata harmonization, and governance oversight, while federated storage and computation can help manage privacy, security, and rights differences across domains and partners. DOE should therefore prioritize interoperable APIs, common metadata schemas, and shared governance standards that allow both centralized discovery and distributed stewardship, including by publishers and specialized repositories.

IP, OTA, and protecting U.S. research and innovation (cross-cutting questions 6–8)

DOE's use of Other Transaction Authority (OTA) under 42 U.S.C. §7256 and related mechanisms offers a valuable opportunity to design modern IP frameworks that both promote AI innovation and respect the rights of content and data providers. STM encourages DOE to use OTA flexibilities to enable tailored data- and content-use arrangements that recognize the value of curated scientific

publications and databases, including paid and open-access models, while avoiding any default assumption that generic or federal purpose clauses override existing copyright and licensing arrangements. As noted earlier, policies that undermine copyright and IP may seem expedient in the short term but undermine American competitiveness in the long term.

STM therefore urges DOE to: (1) clarify with more precise vocabulary and distinguish between copyrighted works and raw data, in any references to "data" or "deliverables"; (2) ensure that IP and data-rights terms in OTAs and cooperative agreements (including CRADAs) are narrowly tailored, respecting pre-existing licenses and contractual arrangements for publisher content; and (3) explicitly recognize that nothing in DOE agreements should be construed to override or nullify private-sector licensing terms for copyrighted works unless the rightsholder has expressly agreed.

Protecting U.S. research and technology security interests—including through DOE's Research, Technology and Economic Security (RTES) requirements—is essential, particularly for frontier AI models with dual-use implications. Because RTES explicitly includes preventing the theft or loss of U.S. intellectual property, strong copyright, licensing, and trade secret protections, and their corollary enforcement, must remain central to DOE's approach. Consistent with these principles, DOE should avoid IP or data-rights provisions that inadvertently discourage private sector investment in high-quality scientific content and infrastructure in support of U.S. competitiveness in scientific publishing and related knowledge industries.

In designing the consortium's legal and organizational structures, DOE should explicitly provide for participation by scholarly publishers and related infrastructure providers to help ensure that IP frameworks, licensing mechanisms, and integrity safeguards are built into the consortium's design from the start, and that any novel IP or data-rights constructs developed are crafted with direct input from those whose content and services are essential to trustworthy AI.

STM and its members stand ready to work with DOE, the National Laboratories, AI developers, and other partners to collaboratively design a consortium that advances transformational AI for science and engineering while safeguarding the legal, ethical, and integrity foundations on which trustworthy research depends. Feel free to reach out to me or to David Weinreich, Director of Policy and Government Relations, with any questions.

Respectfully submitted,

Dr. Caroline Sutton
CEO
STM

Appendix: *Providing feedback or changes to specific terms and conditions or DOE policy*

The following specific suggestions are aimed at promoting the use of high-quality, validated works in AI training through licensing and to ensure that quality and integrity is promoted throughout AI training, development, and deployment in scientific and research settings. They are meant as specific examples in service of STM's overarching recommendations: to embed precise, layered definitions in its program documentation and expressly acknowledge the distinct status of copyrighted content and the role of licensing in enabling lawful and high-quality AI development.

Overall, STM urges DOE to clarify, in any solicitations or consortium agreements, that "data" does not include copyrighted publications and other protected works unless they are explicitly licensed or provided under clear terms that allow their use for AI training, evaluation, or deployment. Any definitional language should distinguish clearly between: (1) government-generated or public-domain data; (2) research data subject to specific contractual, privacy, or security controls; and (3) copyrighted publications and proprietary databases that require explicit permission or licensing for use. In making these recommendations, STM notes that under OTA, DOE is not bound to use the standard Rights in Data Appendix A data clauses; that flexibility can be used to support flexible, innovation-oriented instruments for the complex R&D that is AI.

1. Definition of "Data" and scope of rights

Issue: The RFI and related DOE instruments often use "data" in a broad, undifferentiated way that risks sweeping in copyrighted publications and proprietary databases as if they were unprotected raw data.

Proposed wording (clarification of "Rights in Data" definition aligned with goals):

> "Data" means recorded factual material, including scientific measurements, observations, experimental results, and associated metadata, that are necessary to generate and validate research findings and enable reuse of research outputs. "Data" does not include scholarly publications, books, journal articles, conference proceedings, or other literary or artistic works protected by copyright, nor does it include proprietary databases or content products, except to the extent such materials are expressly identified and licensed for specific uses under this agreement.

Rationale: This clarification ensures that DOE's treatment of "data" aligns with long-standing distinctions in U.S. law between facts and copyright-protected expression, while leaving full room for parties to license use of copyrighted publications and proprietary content for AI training or evaluation where mutually agreed. Absent such clarification, seemingly neutral references to "data" could be interpreted to authorize or encourage broad ingestion of copyrighted works of any kind into AI training pipelines without authorization, which would be inconsistent with U.S. IP law, existing licensing markets, and the Administration's stated commitment to respect for intellectual property in AI policy. Appendix A to Subpart D of 2 CFR Part 910 already distinguishes between categories of data, and this would further reinforce those distinctions.

2. Standard data rights and licensing overrides

Issue: DOE data-rights clauses could be interpreted as granting broad government-purpose licenses over all "data" or "deliverables," including pre-existing copyrighted or proprietary content.

Proposed wording (data rights / deliverables clause):

> "Nothing in this clause shall be construed to grant the Government, laboratories, or third parties any rights in pre-existing copyrighted works, proprietary databases, or other content, beyond those rights expressly granted in a separate license or other written agreement. Rights in such pre-existing materials remain governed by the applicable license terms or contracts."

Rationale: This language preserves the integrity of private licensing markets and encourages rightsholders to contribute high-value content under clear, negotiated terms, instead of deterring participation for fear of inadvertent expropriation through boilerplate clauses. DOE would thereby acknowledge the distinct status of copyrighted content and the role of licensing in enabling lawful and high-quality AI development.

3. Tailored rights in models and training sets

Issue: Many current federal clauses do not distinguish clearly between (i) raw research data, (ii) curated content and corpora assembled for AI training, and (iii) trained models themselves, nor between copyrighted and non-copyrighted materials

Proposed wording (AI-specific rights clauses):

> "The parties acknowledge that: (a) training inputs and corpora may include a mixture of government data, third-party content, and proprietary materials; (b) rights in such inputs shall be allocated in accordance with the rights in the underlying components; and (c) no party shall obtain, by virtue of this agreement alone, any right to use third-party or proprietary content included in training input for purposes other than those expressly authorized under the applicable licenses."
>
> "Rights in models developed under this agreement shall be specified separately from rights in training inputs , with due regard to the protection of proprietary and copyrighted inputs and to applicable national security and export-control requirements."

Rationale: Appendix A allows for special data provisions for particular projects or data sets. Clear separation of rights in inputs, training corpora, and models reduces legal uncertainty, supports investment by content providers, and enables DOE to deploy models widely without unintentionally appropriating underlying content. This is consistent with current practice under 37 CFR 401.14: inventions (potentially embodied in models or model components) are governed by the patent rights clause, but copyright and database rights in training inputs remain governed by copyright law and corollary rights in data, such as their selection and arrangement.

4. Use of OTA and CRADA authority for content licensing

Issue: OTA and CRADA mechanisms provide flexibility but could, without explicit guardrails, be used to impose de facto compulsory licensing terms on copyrighted or proprietary content.

Proposed wording (OTA / CRADA policy statement):

> "When using OTA or CRADA authorities in connection with AI-related projects, DOE and its laboratories will not require participants to grant rights in pre-existing copyrighted works or proprietary databases beyond what is necessary for the performance of the project and will by default rely on market-based licensing arrangements for such materials."

Rationale: OTA explicitly allows DOE to depart from standard assistance and procurement regulations and tailor IP and data terms to project needs. This language signals that DOE will prioritize voluntary, market-based licensing for publisher content and databases, while still retaining flexibility to negotiate specific rights needed for project performance. This is consistent with 2 CFR Part 930's emphasis on flexibility and tailored terms.

5. Transparency on provenance, licensing, and model documentation

Issue: Current DOE terms focus appropriately on security and compliance but could more clearly require documentation of provenance and legal bases for training data.

Proposed wording (added documentation requirement):

> "For any AI model developed or substantially trained under this agreement, the responsible party shall maintain documentation describing: (i) the categories and principal sources of training data; (ii) the legal bases (e.g., public-domain status, license, contract) for the inclusion of non-public or copyrighted material; and (iii) any restrictions on downstream use arising from such legal bases, including license terms. This documentation shall be made available to DOE and consortium partners under appropriate confidentiality protections."

Rationale: 2 CFR Part 930 allows DOE to add project-specific conditions that support compliance and oversight. Requiring concise documentation of sources and legal bases helps ensure legal compliance, supports reproducibility and scientific integrity, and facilitates appropriate downstream licensing without unduly delaying partnerships.

6. Recognition of publisher participation and IP stewardship

Issue: FOA and consortium templates already anticipate industry, academic, and nonprofit partners but do not explicitly recognize publishers as a distinct category, even though their contributions map directly onto expected roles. Current templates do not explicitly anticipate roles for scholarly publishers as core partners, which can lead to misaligned IP expectations and underutilization of their expertise. In designing the consortium's legal and organizational

structures, including potential incorporated consortia, Focused Research Organizations, or other awardee models, DOE should explicitly provide for participation by scholarly publishers and related infrastructure providers as core partners, not only as downstream users.

Proposed wording (participation / roles sections):

> "Consortium membership may include institutions of higher education, scholarly publishers and related infrastructure providers, other non-profit and for-profit organizations, and state and local governments."
>
> "For such partners, DOE will recognize existing IP and licensing frameworks as essential enablers of trustworthy AI and will structure agreements to respect and leverage those frameworks rather than override them."

Rationale: Explicitly recognizing publishers' roles and collective IP stewardship responsibilities helps align expectations, encourages participation, and embeds accuracy and integrity into AI projects from the outset. Scholarly publishers and related infrastructure providers should be seen as core partners responsible for providing high-quality validated and curated content and expertise in the communication of research findings and research integrity, as well as value-added discovery tools and quality metadata. The proposed wording slots into those existing structures to ensure that publisher participation and IP stewardship are explicitly accommodated within DOE's standard practices. This will help ensure that IP frameworks, licensing mechanisms, and integrity safeguards are built into the consortium's design from the start, and that any novel IP or data-rights constructs developed under OTA are crafted with direct input from those whose content and services are essential to trustworthy AI.