August 22, 2019

STM Response to the Request for Information on Identifying Priority Access or Quality Improvements for Federal data and models for Artificial Intelligence Research and Development (R&D), and Testing

The International Association of Scientific, Technical and Medical Publishers (STM) is the leading global trade association for academic and professional publishers. At STM we support our members in their mission to advance research worldwide. Our 150 members based in over 20 countries around the world collectively publish 66% of all journal articles and tens of thousands of monographs and reference works. As academic and professional publishers, learned societies, university presses, start-ups and established players we work together to serve society by developing standards and technology to ensure research is of high quality, trustworthy and easy to access. We promote the contribution that publishers make to innovation, openness and the sharing of knowledge and embrace change to support the growth and sustainability of the research ecosystem. As a common good, we provide data and analysis for all involved in the global activity of research.

The majority of our members are small businesses and not-for-profit organizations, who represent tens of thousands of publishing employees, editors and authors, and other professionals across the United States and world who regularly contribute to the advancement of science, learning, culture and innovation throughout the nation. Our members, comprise the bulk of a $10 billion publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade.

STM supports publishers in their mission to advance research worldwide, including the development of tools that make research more efficient and trustworthy, and the supply of information that can support such development. STM publishers are stakeholders in the development of artificial intelligence (AI), curating and making available scholarly information and data and continually developing new formats for which to do so, including formats that are interoperable with AI and AI products. We therefore support the Administration's efforts in the National Artificial Intelligence Research and Development Strategic Plan and welcome the opportunity for further collaboration with funders and policymakers as it is implemented.

We are encouraged by OMB's efforts to ensure the widest possible use of data the federal government creates in order to drive innovation and new applications. As OMB and federal agencies

work to advance this objective, we appreciate that OMB has recognized that there are data ownership and intellectual property concerns that must be taken into consideration when providing these resources, and encourage OMB to utilize existing frameworks, technical standards, and metadata protocols to ensure the appropriate use of data and the trustworthy development of AI. Appropriate recognition of the investments necessary to create, curate, maintain, and provide metadata for high-quality inputs to machine learning will help ensure the development of high-quality and trustworthy AI.

As agencies review their data and models, what are the most important characteristics they should consider?

With respect to the first set of questions regarding "additional access for data and models," we encourage the Federal government to build on existing technologies and standards to guide any technical requirements with respect to formats, structure, and metadata. Throughout the academic and research community, which includes but is not limited to scholarly publishers, there have been huge strides in enabling and expanding the sharing and use of data. In particular, communities have developed infrastructure, standards, and policies to enable data sharing, including the use of Digital Object Identifiers (DOIs) through CrossRef and DataCite, interlinking protocols developed in frameworks like SCHOLIX, and efforts to improve metadata. These could be harnessed to support the National Artificial Intelligence Research and Development Strategic Plan. The government should utilize existing standards and practices to ensure that its efforts to better leverage Federal government data are aligned and consistent with efforts within the broader data community for non-governmental data. This will ensure that Federal data can be used alongside non-Federal sources in machine learning and the development of AI.

What characteristics should the Federal Government consider to increase a data set or model's utility for AI R&D (*e.g.,* documentation, provenance, metadata)?

It is well known in computer science that the result of a process is only as good as the inputs to that process, often referred to as GIGO, or "garbage in, garbage out." Therefore, it is critically important that any data provided to a machine learning system be well documented. The availability and accessibility of high-quality training data is vital for empowering AI developers with the materials required to achieve both deep learning and to unlock the great potential of AI. STM's members are at the forefront of digital innovation, providing stored and organised information, tagging and enriching content and creating ontologies. All of these advancements, together with the accuracy of the scientific record maintained by science and academic publishers, help to ensure that machine learning has both depth and accuracy. Ensuring that these materials are appropriately tagged to identify the quality metrics and any curation or vetting of the underlying information can only help to make the data more useful and the resulting AI tools more trustworthy.

Likewise, where appropriate, information about licensing or restrictions on use of data should be indicated to provide developers with a secure and legally reliable framework for product

development. This can be provided in metadata or other documentation. The wide array of licenses offered by publishers for the material they currently provide for machine learning and AI developers ensures that there are ample, accessible materials available for the continued training of both people *and* machines.

What data ownership, intellectual property, or data sharing considerations should be included in federally-funded agreements (including, but not limited to, federal contracts and grants) that results in production of data for R&D?

A careful balance needs to be struck between the government's interest in data collected pursuant to Federal contracts and grants, and incentives for those in the private sector to work with the government to also collect and improve data for R&D. The need for balance is enshrined in Federal R&D policy, most clearly in the Bayh-Dole Act of 1980 but also within the provisions of the OPEN Government Data Act and the Federal Data Strategy, all of which call for the recognition and safeguarding of private sector intellectual property rights.

The priority for improved access to data under the AI strategy should be "programmatic, statistical, and mission-support data" that originates with, and therefore belongs to, the Federal government. Where contractors are responsible for collecting such data within the scope of work of a contract, requirements for government ownership of the resulting data might be appropriate in some cases, but not in all. Under grants, however, where research may be open ended and the project design and implementation remains the intellectual property of the awardee, it is more appropriate for the grantee to be able to determine the conditions under which the data can be used. Without providing such rights to the awardee, the government risks undermining the market mechanisms that encourage discovery and innovation, and provide the incentives for tagging, formatting, and structuring data. This could counter the goals of the National Artificial Intelligence Research and Development Strategic Plan to enable broad access to, and use of, such data for AI and R&D, as well as undermine the economic benefits that are derived from private sector investment and the commercialization of these technologies.

The Federal government should be clear that information and works that are subject to private sector intellectual property rights, such as journal articles, books, or proprietary datasets, will not be subject to dissemination or other requirements with respect to AI R&D so that these ancillary products are incentivized to the full extent envisioned by our copyright and patent systems. As partners in AI development, STM publishers ask that the information products they and others bring to the development of AI, and the consequent value that they add, be appropriately recognised in the future development of AI products that use this information and data.

The International Association of Scientific, Technical and Medical Publishers (STM) supports further collaborative work to ensure that all stakeholders benefit from the great potential of AI. Publishers are already meeting the needs of the AI era, by developing tools, services and platforms that support and enhance machine learning. STM's members are at the forefront of digital innovation, providing

stored and organised information, tagging and enriching content and creating ontologies. More information on our efforts is available in our 2018 statement, *STM Publishers Innovations Support AI and Machine Learning.*[1] We look forward to continued engagement with OMB and the Administration towards these goals.

Sincerely,

Michael Mabe
CEO STM

[1] https://www.stm-assoc.org/2018_10_08_STM_Publishers_Innovations_Support_AI_and_Machine_Learning_8Oct2018docx.pdf