CONVENINGS **#2025-01**

Expanding Access to High-Quality, Factual, Philanthropic Data

Version 1.0

Rapporteur
Katherine Townsend
Open Data Collectives

Conveners
Andrew Gruen and Bennett Hillenbrand
Working Paper

11 November 2025

Working Paper
workingpaper.co

1 Introduction

As AI tools permeate enterprises, governments, and daily life, social impact organizations face a dual challenge: adopting new technologies while protecting under-resourced and marginalized communities from harm. Philanthropies and research organizations navigate troves of narrative reports and decades of individual expertise and shifting metrics to inform changes to complex systems from migration to vaccine hesitancy to democratic resilience. Large language models provide support, but the emergence of the Model Context Protocol (MCP) significantly expands the potential for reliable, source-grounded analysis that supports expert decision-making in the social sector.

Working Paper recently convened a Chatham House–style gathering of individuals from fifteen organizations to share early experiences building and applying MCPs for social impact organizations. This document summarizes key insights, participant discussions, and major themes from the presentations, and highlights implications for the field.

2 Key Insights

- MCPs enable reliable insight generation from trusted, authoritative data without needing to completely disclose the underlying training data.
- When paired with LLMs, MCPs allow queries of complex datasets in natural language without deep data-analysis skills, which expands opportunity and insights to novices, but can give a false sense of truly deep expertise.
- MCPs help surface insight from decades of unstructured and semi-structured narrative reporting.
- MCPs may reduce compute, carbon, and water usage by turning some answers into lookup queries instead of token generation tasks—in addition to storing and reusing validated answers rather than regenerating them for each query.
- High-quality data remains the primary driver of training and using accurate models, and is chronically under-funded due to the intensive work required to produce it and the challenge with tying that work to realized returns.
- Adversarial actors understand this high-quality data gap and flood training corpora with disinformation, including state propaganda that appears in globally popular LLMs. The implications are particularly problematic for decision-makers that look to LLMs alone to inform understanding.

3 Discussions

Participants compared methods for ensuring accuracy in AI models and sought clarity on how to guarantee that commercial LLMs and agents use validated MCP-based results rather than bypassing them. Most agreed that this will require formal partnerships and technical safeguards.

Conversations about data accuracy centered on hallucinations and the challenge of new, unvetted data entering models. There was less discussion of trustworthy public data sources, though participants noted ongoing changes to U.S. government datasets. Historic datasets maintained by participating organizations were unchallenged as high quality.

Data repository hosts highlighted the difficulty of providing secure, intuitive navigation and remote analysis tools for researchers who would prefer to download data locally, which is an impossible task when datasets span petabytes. They also discussed the tension between investing in user-friendly tools as a genuine pathway to solutions discovery versus casting a wide net as a marketing function. This raised an existential question about who the “users” of academic data sets have been in the past versus who they can or should become in the modern AI era.

Funders and grantees acknowledged persistent gaps in funding for compute, storage, infrastructure, bandwidth, and legal costs, highlighting a broader misunderstanding of what modern data work requires. One example involved a grantee in rural DRC working in an unmapped French dialect, who bore unexpected additional costs as a result of a requirement they were given by a funder to use specific AI tools.

Participants also raised concerns about a growing digital public goods problem: organizations with resources are currently focused on building custom AI integrations that satisfy their primary stakeholders. These efforts often prioritize serving their narrow internal needs rather than contributing to shared infrastructure for the field.

4 Presentations

4.1 ICPSR: The Need to Expand Capacity for Solution-Seekers to Find the Right Data

Libby Hemphill, Inter-university Consortium for Political and Social Research

Dr. Libby Hemphill described ICPSR’s vast data repository and the challenge of designing an interface that is useful to users of varying degrees of expertise. She outlined three user pathways: experts who navigate “directly,” novices who take the “scenic route,” and semi-expert users who remain stuck in an “orienting” phase. ICPSR’s metadata knowledge graph, coupled with MCPs, allows LLMs to surface relevant datasets even when user queries are vague or poorly aligned with conventional search methods. This feature is designed for those on the “orienting path”: semi-experts who know the domain and the problem, but not how to translate that knowledge into technical data queries or identify relevant datasets. This change should expand access to the creation of data-driven solutions. Dr. Hemphill also emphasized that access to data is often fragmented by sensitivity, contractual limits, and compute constraints, underscoring the need for mediated environments for sensitive data.

4.2 Ensuring Accurate, Authoritative Representation of Data with MCPs and PCNs

Aivin Solatorio, World Bank

Aivin Solatorio focused on the need for accuracy, verification, and trust in multilingual environments and the implications for the billions impacted by World Bank's policies. Many of the Bank's decisions can be understood and may be driven out of information found on its open data portal, Data360, which receives over 100 million visits annually, with 30% of queries in languages other than English. The Bank expects chatbots, including those created by users, to be primary access points for its data. However, even Retrieval-Augmented Generation (RAG) systems built to ensure accuracy can hallucinate when pressured to produce an answer. To address this, the Bank is building an MCP layer paired with Proof-Carrying Numbers (PCN), which require AI responses to include cryptographically verifiable references to authoritative data. If the model cannot regenerate a fact according to source values, the system rejects the output.

4.3 Building AI Commons with Multimodal Embeddings and QAG

Robert Grossman, University of Chicago

Dr. Grossman outlined his work building The Genomic Data Commons (GDC), the world's largest collection of harmonized cancer data, currently holding 60–70 petabytes and having been cited in more than 30,000 papers. Grossman shows that the architecture required to realize the future that MCP+LLMs are proposing already exists at scale: data commons, multimodal embeddings, small models, and RAG/Query-Augmented Generation (QAG) pipelines. He effectively argues that AI only works on top of high-quality, well-organized data. The future is not "search engines over data"; rather it's AI systems that sit on curated, multimodal embeddings that let users build or run real models on high-density information, rather than relying on, and being limited by, natural language queries.

Grossman proposes expanding his experience of building the GDC, and generalizing it through Gen3, an open source tool for others to develop their own data commons, to now architect what he calls AI Commons. To ensure AI accuracy, his team developed QAG, a pipeline where an LLM extracts intent, generates precise API calls to the data commons, retrieves real results, and checks that the final answer hasn't been hallucinated. As an example, when seeking specific gene frequency data, QAG achieves 0% error, compared with significant errors from frontier models.

The foundation of the AI Commons approach is multimodal embeddings. He shares the example that with high-quality embeddings built from RNA-seq, histology, clinical notes, and other modalities, small and mid-scale models routinely outperform giant frontier models. Grossman demonstrated that simple statistical modeling techniques (like CoxPH) paired with curated, high-quality multimodal embeddings outperform state-of-the-art AI published in major scientific journals. Knowing a patient's age and cancer type already yields ∼0.73 concordance; adding doctor notes and embeddings raises this to ∼0.79—better than "big model" results.

He outlines three key tensions:

1. **Data Quality.** Citing scaling laws, Grossman argues that data quality is the real bottleneck in biomedical AI, not model size.

2. **Interactions with data need to be simple.** AI Commons need to be secure, flawlessly navigable, and any AI agents employed need to be grounded in the data.

3. **Moving beyond sharing datasets to sharing models.** AI Commons must support: embeddings by default; small and mid-scale models built by users; auditability and reproducibility; and actual usage, not static downloads.

His proposed future is an embedding-centric AI architecture—not trillion-parameter models, but compact, accurate systems built on curated multimodal data, running on a few GPUs, and accessible to any researcher. This is the backbone of the University of Chicago's Meshes of Midscale Models (M3) initiative. This proposed approach substantially reduces the costs and compute required to derive high-quality scientific insights and findings.

4.4 Scalable, Operational Infrastructure in the Commercial World

Dave Hotchkiss, Google Data Commons

Google's Data Commons integrates public data from national statistical offices, NGOs, and other sources, often in heterogeneous formats such as PDFs, and makes it usable for policymakers, analysts, and researchers. It also powers Google Search "knowledge panels," with transparent source links. Google's Retrieval-based Inference Grounding (RIG) annotates AI responses with authoritative data points and shows measurable improvements over RAG alone. Hotchkiss noted that Google now supports private data-commons instances that can federate with the global commons, enabling large-scale use while respecting licensing and sovereignty constraints.

Countries are quickly building out their AI infrastructure and capacity and many are recognizing the need for high-quality data to inform accurate insights. Hotchkiss acknowledged the limitations of publicly available authoritative data and the benefit of working with countries to expand their open data repositories to build trust, attract investors, and improve AI-driven analysis. He cited interest from several countries, including small island nations, that are proactively improving their data to better engage with AI tools. Hotchkiss described a partnership with the ONE Campaign to test Google's MCP for policy use. In follow-up conversations, ONE's data team reported that MCP queries against World Health Organization data were accurate, but cautioned against over-reliance on MCP-generated policy language given persistent gaps in data from non-English-speaking countries and global-majority communities, and the broader risk of algorithms supplanting expertise.

4.5 How Social Impact Organizations Can Implement MCPs and RAG Effectively

Sam Patel & Lucky Gunasekara, Robert Wood Johnson Foundation & Miso.ai

For philanthropies, and many organizations with decades of historic data, information is not structured for AI and does not show up or shows up poorly in LLMs, resulting in inaccurate representations, hallucinations, and vulnerability to mis/disinformation. RWJF sought to adopt a stack to better understand the potential of MCPs to historical information and improve agentic tools used by employees and external audiences alike.

Their demonstration showed that an MCP-enabled agent can: search internal grant history; recall insights from past research; answer policy questions; route users to grants or jobs; and provide consistent, non-hallucinated summaries.

RWJF was able to achieve these results in large part due to the fact that they already had a well-structured historical dataset. While any large mission-driven organization can replicate this by structuring its data, this is no small task. The best time to start this—as Patel showed—was decades ago. But the next best time to make an organization's data machine-readable, so it can be connected through MCP/RAG layers, is now.

4.6 Propaganda is Already Influencing Large Language Models

Solomon Messing, Center for Social Media and Politics, NYU
with Hannah Waight, Eddie Yang, Yin Yuan, Margaret Roberts, Brandon Stewart, and Joshua Tucker

Dr. Messing presented research from his team examining how political systems, and in particular authoritarian regimes, shape the training data that large language models absorb. This research shows that LLMs trained on broad internet content inevitably ingest state media, which in turn shapes how frontier models respond to prompts across many languages. While many findings focus on China's Communist Party, the pattern holds across 37 countries and thus far, frontier models have not addressed the issue.

By influencing what's on the web, China and other authoritarian regimes may shape understandings of modern geopolitics for the growing populace who get their information from LLMs and agentic AI. The potential for strategic manipulation may present itself, if LLM developers do not pay more careful attention to the mix of their training data and its downstream effects.

Key takeaways:

- There is no neutral training corpus.
- Authoritarian states exploit the training pipeline (whether intentional or not).
- All major LLM developers examined in the research failed to adequately filter state media sources.
- Post-training guardrails are so far insufficient when propaganda is baked in at pretraining.
- Greater transparency in model training procedures—pretraining in particular—could help to address these concerns.

5 Conclusion

MCP adoption is accelerating across the social sector, giving organizations a real path to generate better insights from trusted data. In many cases, nonprofits and research institutions are outpacing the private sector in building reliable, natural-language access to high-quality datasets. Participants agreed: MCPs dramatically outperform standalone LLMs, especially when paired with safeguards like PCN, QAG, or well-designed RAG systems.

We were encouraged by tools that make complex data usable for experts without technical training, and cautiously optimistic that MCPs may improve the quality and reliability of AI-generated answers while reducing the compute demands of generating those answers.

The risks we currently bear with generative AI are substantial but not insurmountable. Disinformation and state propaganda is already baked into popular LLMs. MCP+LLMs can make weak or incomplete data appear authoritative, accelerating policy or investment decisions that overlook context, gaps, or marginalized voices. Chronic underinvestment in high-quality data pipelines, linguistic inclusion, and compute only compounds this vulnerability.

As organizations build their own MCP layers, we assert that the fundamentals still hold: good data requires expertise and investment; interoperability matters; and adversaries will exploit every weakness. Use these tools, but verify their insights. We will be looking eagerly for the following evidentiary signals of high-quality implementation:

1. Transparency around investment in data engineering and confidence in data quality and curation.

2. Following the World Bank's lead in PCN or similar efforts that greatly increase the reliability and trustworthiness of the results from GenAI systems—regardless of the interface method.

3. Leveraging MCP or similar technologies to make search interfaces public, interoperable, and open source so a true global data and AI commons can emerge as a digital public good.