APPENDIX B: TAXONOMY OF DATA COMMONS USE CASES FOR AI

To supplement The Open Data Policy Lab's Blueprint to Unlock New Data Commons for AI, this document provides two frameworks that organizations that have or steward data can use to advance new data commons to enable public-interest AI use cases. This includes, for instance, knowledge institutions (such as libraries, universities, research centers) and others that hold unique and high-quality information assets (such as governments, non-profits).

The first framework maps the types of data that can be useful for generative AI applications. The second framework outlines 10 distinct AI use cases where a data commons approach could be valuable for society. These frameworks are a starting point only and aim to be exemplary, not all inclusive. We hope these resources inspire new data commons initiatives that help unlock AI's full potential.

CONTEXT

To organize the energy around data commons in a productive fashion, The Open Data Policy Lab conducted a series of structured analytic exercises to ideate potential use cases.

This effort began with a mapping exercise, identifying sectors where data commons for AI already exist and where there was substantial interest in securing more access to data. We relied on a search on the web (e.g. Google Scholar, news outlets, The Open Data Policy Lab's *Observatory of Generative AI and Open Data Use Cases*, and The GovLab's *Data Collaboratives Explorer*) and our network of experts (e.g. Open Data Action Labs). This review helped us understand current efforts across industries. We summarized this work within the blog: *The Emergent Landscape of Data Commons: A Brief Survey and Comparison of Existing Initiatives*.

We then conducted another mapping of data types that could be useful for different aspects of generative AI, drawing from Microsoft's *Government Data Commons: A Technical Guide*, submissions to the *United States Department of Commerce Request for Information on AI-Ready Open Government Data Assets*, The Data Provenance Initiative's *A Large Scale Audit of Dataset Licensing & Attribution in AI*, the Open Data Institute's *Data for AI Taxonomy*, *Papers with Code*, and our network of experts.

We focused on data types that were cited as valuable for generative AI or where authors expressed there is a need for data. We chose not to include data types with clear ethical and data responsibility issues (data types such as electronic health records that have been widely criticized by the public for its use within AI systems). We also tried to avoid data types where extensive data commons already exist to avoid the duplication of efforts. Using this information, we developed a taxonomy to organize these data types.

At the same time, we conducted a "What if?" Analysis in which we asked ourselves to consider the types of use cases where new AI data commons could benefit society. Drawing on the information gathered thus far, we synthesized 10 categories of use cases for new data commons in the context of AI.

In what follows we provide a summary of these exercises. While these exercises primarily focus on generative AI, it is important to note that there is still a need for data commons for machine learning and neutral networks as well.

TAXONOMY OF DATA TYPES

Below we provide a list of data types that could be valuable for data commons for AI. These data types are divided into six main categories. Creating data commons that are truly fair and beneficial to all requires sourcing data responsibly. That being said, the data commons should only include data sources with appropriate licensing.

Category	Data Types (Exemplary)
1. Text	• Language and linguistics data: Language research, words, definitions, translations, and transcripts • Legal data: Laws, legal proceedings, and court case decisions • Research and education data: Exam questions and answers, educational blogs, homework questions and solutions, research reports, peer-reviewed journal articles (focusing on niche scientific subjects), grants, and patents • News and media data: Press releases, blogs, newspapers and magazines, scripts, books (e.g. textbooks, cultural perspectives, encyclopedias, travel guides, and translations) • Biomedical data: Notes and insurance policies • Social media data: Social media posts, customer reviews, online petitions, search terms/keywords
2. Audio	• Sounds: Environmental sounds (urban and nature) and physiological sounds (e.g. breathing patterns) • Music • Speech Data: Conversations, expert testimonies, phrases, government proceedings (e.g. policy debates, consultations), storytelling data (e.g. audiobooks, recent radio recordings)
3. Images	• Knowledge graphs • Spatial images: GIS/satellite data, maps • Artwork (including photographs) • Data visualizations: Charts, tables, infographics
4. Video	• Expert lectures • Government videos: Policy debates • Environmental videos: Urban videos (e.g. driving scenarios, traffic camera footage) and nature videos

5. Statistics	• National statistics • Sensor networks • Supply chain data: Manufacturing metrics • Energy consumption data • Math/physics proofs: 19th century physics proofs
6. Generative AI data	• Synthetically created datasets • Data about generative AI providers: Models (weights, usage, performance), development processes • Issues associated with generative AI: Unwanted outcomes, legal cases, other risks • Best practices when using generative AI: Prompts, policies, strategies, and guidelines

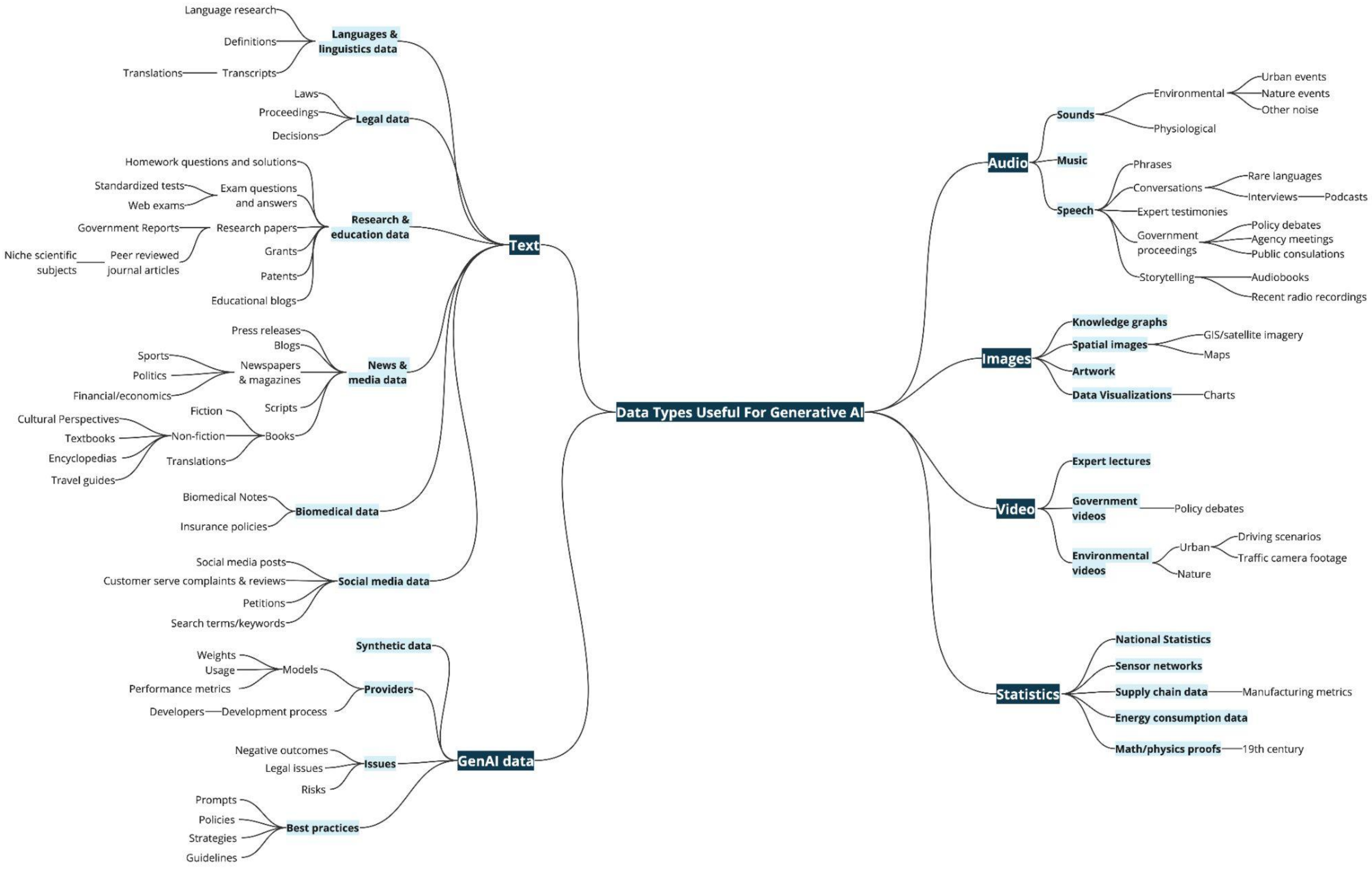

Figure 2. Summary of Data Types

View the full map HERE

TAXONOMY OF USE CASES

The following chart summarizes our taxonomy of data commons use cases that could benefit society. In this taxonomy, we focus on the purpose for developing AI technologies, the types of technologies that could be developed and the scope of the use or who it is affecting most. We also outline potential stakeholders that may be involved. This list of stakeholders is exemplary only and non-exhaustive.

Area of Focus	Purpose	Scope	Potential Outcomes	Stakeholders
1. Research and Scientific Discovery	Enhance scientific exploration and accelerate new discoveries and innovations	Societal	Drug discovery, hypothesis generation, climate modeling, open science collaboration	• Scientists and researchers • Research and academic institutions • Research funders • Publishers/journals • Government research agencies • Multilateral organizations • Representatives of the public
2. Education, Training, and Learning Support	Support education and learning through personalized and interactive tools	Individual	Personalized learning platforms, training simulations, assistive tools, content generation	• Students • Educators • Research and academic institutions • Education technology companies • Government education departments • Multilateral organizations

3. Content Development and Knowledge Preservation	Create and tailor content for education, service delivery, communications, and the preservation of knowledge across cultures	Organizational	Multilingual content, public health campaigns, digital heritage, AI-assisted infographics/visual communications	• Archives, museums, and libraries • Government communications departments • Civil society organizations (NGOs, advocacy groups, and unions) • Multilateral organizations • Representatives of the public
4. Modeling, Simulation, and Anticipatory Decision-Making	Anticipate future outcomes through realistic, data-driven scenario generation	Societal	Climate simulations, policy modeling, urban planning (e.g. digital twins of cities), crisis management, multi-stakeholder decision models, risk analysis	• Foresight and forecasting practitioners • Government strategists and policy makers • Local governments • Philanthropies • Climate agencies • Representatives of the public
5. Pattern Recognition, Inference, and Insight Generation	Analyze and identify trends and anomalies in data for better decision-making	Societal	Traffic optimization, crime pattern simulation, disinformation detection, economic trends	• Research and academic institutions • Government agencies • National Statistics Offices • Local governments • Multilateral organizations • Civil society organizations

6. Personalization and Customization of Services	Tailor services and experiences to individual needs	Individual	Personalized healthcare, adaptive public services, civic engagement tools	• Domain experts • Government agencies • DEI departments • Civic tech • NGOs and multilateral organizations
7. Automation of Repetitive Tasks	Increase efficiency by automating time-consuming processes	Organizational	Automated reporting, legal document drafting, customer support chatbots	• Government agencies • Multilateral organizations • NGOs and non-profits • Philanthropies • Advocacy groups • Businesses
8. Synthetic Data Generation	Preserve privacy while creating synthetic data for model training and research	Organizational	Synthetic datasets for healthcare, research benchmarks, addressing data scarcity	• PII departments • National Statistics Offices • Multilateral organizations
9. Real-Time Insights and Adaptive Systems	Provide feedback in real-time for dynamic system improvement	Societal	Real-time traffic and health monitoring, adaptive disaster response	• Government agencies • Multilateral organizations • Emergency response departments • Local governments

10. Collaborative Intelligence	Enhance collective problem-solving and finding consensus among diverse groups	Organizational	Mediation and conflict resolution, administrative support, public consultation tools	• Government agencies • Multilateral organizations and NGOs • Policy makers • Civil society organizations • Representatives of the public