Government Data Commons: A Technical Guide

A guide for governments to establish high-quality, useful, and beneficial data commons

Microsoft

Government Data Commons: A Technical Guide

Microsoft believes access to data should be equally accessible to developers and innovators in every geography and sector, including startups, open-source developers, and AI researchers. The establishment of open data commons can help us acheive this goal. This guide walks governments through the considerations for preparing data that can form the basis of high-quality, useful, and beneficial data commons.

Data accesibility

Governments should prioritize making public records available online, even with minimal metadata or formatting, to support AI research and development. This will range from providing access to digital records to taking steps to digitize and make available analog content.

Data discoverability

Despite accessibility, data may remain undiscoverable due to web crawling restrictions, format complexity, and poor metadata. Governments should enhance searchability by improving indexing, metadata, and leveraging AI for better discovery.

Data usability

Government datasets can be enhanced by preparing them for AI training. This can include cleaning, annotation, and transformation into formats such as XML. Governments should maintain native formats, use standard formatting, ensure transparency in data provenance, and improve website design to aid AI research and development.

Metadata and data standards

Comprehensive metadata can help with data understanding, context, and usability. Governments should also use open licenses, adopt standards, and enhance discoverability through means such as well-documented data stores and APIs.

Identifying data valuable for AI

Governments should catalogue public data that is impactful for AI development. In the era of generative AI, this includes unstructured data from text, images, audio, and video. Authoritative, unique, knowledge dense, recent, and voluminous data is most valuable for generative AI training.

Tools to facilitate open data

Microsoft's tools and services enhance data accessibility, security, and governance for governments. Key offerings include Azure Digital Twins, Azure Data Factory, Azure Databricks, Azure Synapse Analytics, and Microsoft Purview. Open standards and language model management tools are also valuable to consider.

Microsoft

Introduction

Developing safe, inclusive, and beneficial generative AI models is integral to innovation, economic growth, and human development across geographies and industries. Developing these models effectively, in turn, depends heavily on access to data at scale – this includes a broad diversity of ethically sourced data that are representative of different cultures and languages.

Across countries, languages, communities, and domains, however, there is a lack of accessible generative AI data. This lack of accessibility hinders both AI developers and key populations of users from being able to easily participate in the AI revolution.

Microsoft believes that access to data should be inclusive and equally accessible to developers and innovators in every geography and sector, including startups, open-source developers, and the broader community of AI researchers. This can be achieved through building open data commons. Data commons are pools of data that are managed by communities or entities operating in the public's interest. Data commons offer an alternative to the growing tendency toward privatized data, emphasizing instead the shared value of data.

It is important to anchor the discussion on access to data to how generative AI uses training data. For traditional AI, training data has often included structured data, such as spreadsheets. By contrast, generative AI has a primary need for unstructured data; this includes text, audio, images, and video. The process of pre-training involves introducing a vast corpus of data into the training process. In doing so, generative AI models do not store their training data or use the training data to perform its task. Nor do they search and return items from the training data to provide a response. Instead, generative AI models are designed to generate *new* expressive works and content, not to reproduce content already created.

Generative AI models have the potential to unlock groundbreaking discoveries across sectors, disciplines, and geographies. For governments, nonprofits, and multi-lateral organizations around the world, opening data can help lead to more representative AI and help citizens and government employees leverage AI to improve government responsiveness and efficiency.

Considerations for establishing open data commons include assessing:

- Data accessibility
- Data discoverability
- Data usability
- Metadata and data standards
- Identifying data valuable for AI
- Tools to facilitate open data access

Each of these areas will be addressed in this practical guide for building and establishing high-quality, useful, and beneficial data commons.

Data accessibility

The lack of data accessibility is often the result of several issues:

- **Analog Data:** Analog data is data not stored in a digital format. It includes data stored as paper records, video stored on film, magnetic tapes, and other archival materials. Examples of such data in government records include many national archival records and many books in national libraries. On top of the concern that these formats are fragile and prone to degradation, these formats present many challenges in the digital age that complicates use by multiple parties at scale. Because of its inherently limited physical reach, analog data is the hardest to access.
- **Offline Digital Records:** Offline digital records include digital records stored off-network, within government intranets, or online but behind a secured access database. Examples include large unstructured datasets, such as audio recordings of agency proceedings, which were previously undervalued but are now valuable in the age of generative AI. Government agencies possess vast amounts of offline digital records. Unless AI researchers are granted access under specific circumstances, this data is also inaccessible.
- **Online Digital Records:** Online digital records include digital data that is accessible to anyone with an internet connection. While much of this data is available for AI researchers, such as online PDF databases, other categories are challenging to access due to formatting issues, poor or non-existent metadata, or difficulties in downloading. For example, public domain streaming videos available via embedded media players are hard for AI researchers to download because the files are not easy to locate.

To support accessibility of government data, we recommend establishing a “low bar” for initial accessibility. Governments should prioritize making unstructured data accessible by default, including by making data available on the Internet for download in its native format. This should include all text, audio, images, and videos that are categorized as public records by a government agency or public office.

Given limited resources, the highest return on investment may be for the government to focus on data assets that are already digital but not available online. A second priority could be data assets that are available online but not easily discoverable or accessible. A final focus area could be analog data: data not currently digitized but available for digitization, such as documents in public offline archives that could be scanned.

We strongly believe that data should be made accessible, even if this means simply making it available on the Internet in its native form under an open license, such as the Community Data License Agreement Version 2.0 hosted by The Linux Foundation. Data should not be held back from accessibility to first improve its other qualities, such as metadata or formatting. While we advocate for these improvements, they should not block or delay the data’s baseline accessibility. More valuable datasets can then be made available through more structured formats.

Data discoverability

Even when data is technically accessible, it may not be easily discoverable for a variety of reasons:

- **Web Crawling Restrictions:** Some government websites may restrict web crawlers, preventing search engines from indexing and discovering their content.
- **Format Complexity:** Data stored in complex file formats like PDFs can be difficult for search engines to index effectively.
- **Dynamic and Restricted Content:** Datasets hidden behind search forms, login barriers, or dynamic content are less visible to search engines.
- **Metadata and Schema Deficiencies:** Inconsistent or inadequate metadata and a lack of standardization hinder search engine indexing and data integration.
- **Indexing and Organizational Challenges:** Poor indexing practices and scattered data sources complicate search and retrieval efforts.
- **Search and Filter Limitations:** Databases often lack the ability to filter or sort data by important attributes, making it harder to find relevant information.
- **Lack of Standardization:** Different formats, terminologies, and categorization methods can hinder the ability to find and integrate data from multiple sources.

To support data discoverability, we suggest making data available in popular web indexes, which means ensuring no steps are taken to block crawlers; we suggest investigating the controls associated with your site. It is also important to ensure data is readable by web crawlers for search indexing. Challenges in this context include discovering content behind a log-in or CAPTCHA, rich JavaScript, poor indexing, specialized online forums, and video content that is not indexed.

In addition, glossary terms (aka data dictionaries) can clearly define data fields, units, and relationships to help understand the structure and meaning of data. API documentation can provide clear guidelines for usage and integration to help developers work with the data programmatically. For governments databases that are the largest and most valuable for AI, it will also be useful to focus on better metadata, filtering, indexing, and search capabilities. Data within larger or more complex datasets is harder to discover without the aid of such tools. This is particularly true for very large datasets. We also suggest government agencies continue to develop curated landing pages with links to their most valuable datasets for AI.

It is also worth exploring how AI itself can help users discover relevant datasets within government websites. AI has the capability to understand context across datasets and enable natural language conversations in which a user can ask a question (*e.g.*, “What data is relevant for training an AI model on sign language recognition?”) and the AI experience can return a set of links to databases relevant to the user’s question. Examining how to leverage AI to improve discoverability holds significant promise.

Data usability

Even when data is both accessible and discoverable, it often needs significant preparation to be useful:

- **Formatting and Standardization:** Data from different sources may need to be converted into a consistent format and structure to be used effectively.
- **Data Cleaning:** Raw data may contain errors, inconsistencies, or irrelevant information that needs to be corrected or removed.
- **Annotation and Labeling:** Data, especially unstructured data like text, images, or audio, often needs to be annotated or labeled to be useful for machine learning and AI applications.
- **Integration:** Combining data from various sources can be challenging due to differences in formats, scales, and reference systems.
- **Scalability:** Handling large volumes of data requires scalable storage and processing solutions, which may not always be available or cost-effective.

To facilitate the usability of government data, we suggest maintaining access to data in its native format so that the source material (source of truth) always remains available. This could be PDFs for rich documents, for example. Although AI researchers may often transform these documents into other formats such as JSON and XML to facilitate AI training, a data-providing institution need not do so. Technology is swiftly advancing to make transforming PDFs easier via AI itself, and the most valuable and useable long-term format will not necessarily be easy to predict. If, however, there are very high-value datasets seeing large demand, it could make sense to expend resources to develop a JSON or XML version of a rich document to make available alongside its native format.

It is also highly useful to provide a variety of data types in standard formats and make them scriptable. Universally accepted formats, such as CSV and text files, are commonly used for data dissemination. They are lightweight, easy to parse, and widely supported across tools and platforms making them ideal for efficient AI algorithm data processing. Machine-readable formats including CSV or JSON are easy to download and reuse and offer APIs to ease data access and integration.

In addition, enhancing the design and structure of websites can make data more accessible to LLMs. This includes using structured data markup, keeping website content up-to-date, and maintaining proper security certificates.

Governments should also help AI researchers facilitate the download of the vast amounts of data available. This could include providing easy downloads of a full dataset or a filtered dataset via standard download means or APIs. Often, this is most challenging for video. Many websites provide streaming video or audio access designed for human consumption—which makes downloading data for AI research challenging.

Transparency in data provenance and collection methods is also critical. It can be particularly useful in assessing potential biases in data. In highly consequential domains, such as medical or criminal justice contexts, these considerations take on critical concern. Recognizing that there may sometimes be a trade-off between making data available immediately and waiting for bias analysis, data provenance and potential bias information should be noted and disclosed as much as possible. Important elements of such transparency could include making data available with tags that show whether bias analysis was conducted or not. More broadly, governments can enhance usability by including detailed metadata about data sources, context, and quality. This approach helps users understand the limitations and usability of the data.

Metadata and data standards

It is important to have comprehensive metadata, including detailed information about data sources, context, and quality. It is crucial for understanding the data and its limitations. With semantic metadata, controlled vocabularies or ontologies can greatly enhance the understanding of data. This type of metadata makes it easier for AI systems to interpret the context and relationships within the data. Additionally, metadata can serve as the link between documents and other dataset components. While tagging individual documents with metadata is useful, it is also important to capture the relationships between documents to provide better context and usability.

Understanding the distinctions between raw data and derived data plays a crucial role in effectively applying metadata standards. Raw data, such as data collected from sensor networks, and derived data, such as statistical data, each present unique characteristics and requirements for metadata.

Raw data is unprocessed, voluminous, and granular, often captured in real-time. Making raw data useful requires significant processing and proper metadata standards:

- **Descriptive Metadata:** Details about the data source, such as sensor type, location, collection time, and format, to understand context and provenance.
- **Structural Metadata:** Information on the data format and structure, such as file formats and schemas, crucial for data parsing and processing.
- **Administrative Metadata:** Details about data ownership, access rights, and governance policies to manage and access the data appropriately.
- **Real-time Metadata:** Covers data timeliness, update frequency, and synchronization with other sources.

Derived data, on the other hand, is processed and analyzed data. Metadata standards for derived data include:

- **Descriptive Metadata:** Explanations of methodologies, purpose of data derivation, and context of use to understand how the data was created.
- **Provenance Metadata:** Tracks the origin, processing steps, algorithms, and transformations applied to support transparency and reproducibility.
- **Quality Metadata:** Includes quality metrics such as accuracy, completeness, and reliability to assess the trustworthiness of the data.
- **Usage Metadata:** Information on how to use the data, including limitations and recommended applications, to provide a strong scope of understanding.

To support broad, equitable, and open access to datasets and metadata, Microsoft advises that governments, to the extent a copyright exists, adopt several key practices including open and permissive licenses. We recommend adoption of the Community Data License Agreement, hosted by the Linux Foundation, found here. Other relevant options are the license templates presented by Creative Commons (CC BY) that allow free use and modification with/without attribution. Key elements to consider with any license are to ensure: (a) the terms are clear and accessible; and (b) the license is compatible with other open data licenses to facilitate integration.

With respect to data standards, Microsoft recommends accounting for the following considerations:

- **Current Standards:** The Data Catalog Vocabulary (DCAT), developed by the W3C, describes datasets to make them easily discoverable and usable by AI systems. It facilitates efficient data sharing and integration by publishing metadata about datasets. Schema.org structures metadata in a standardized way, enhancing web crawlability and making data more accessible for AI applications. The OpenAPI standard specifies how to describe, produce, consume, and visualize RESTful web services, making it easier for developers to access and integrate data via APIs.
- **Emerging Standards:** The World Wide Web Consortium (W3C) Data on the Web Best Practices are evolving guidelines aimed at improving the usability and interoperability of data published on the web, supporting better data sharing practices for AI systems. The ISO/IEC JTC 1/SC 42 international standard focuses on AI, including guidelines for AI-specific metadata and data governance, to support AI system responsibility.
- **Enhancing Discoverability:** Governments can develop well-documented APIs to allow seamless data retrieval and integration into AI models. These APIs should support high data throughput and scalability to effectively handle large datasets. Governments can also optimize data repositories for web crawlability by implementing structured metadata and adhering to standards to improve discoverability by AI systems. By structuring data in readable and digestible formats such as CSV and JSON, compatibility with various AI tools and frameworks is made easier.
- **Comprehensive Data Standards:** Microsoft suggests implementing comprehensive metadata standards such as Data Catalog Vocabulary (DCAT) or Federal Geographic Data Committee (FGDC) so that datasets are well-documented and easily discoverable. Interoperability standards, like OGC for geospatial data, support system compatibility. Establishing data quality standards, including validation protocols and regular updates, can maintain high data quality.

Identifying data valuable for AI

Given that all organizations have limited resources, below we offer a framework on how to think about data value prioritization for generative AI, along with a few suggestions on types of datasets to target for accessibility, discoverability, and usability.

For text data for LLM training, Microsoft researchers are finding that data that is most valuable is authoritative, unique, knowledge- and/or logic-dense, recent (or still relevant), and long-context.

- **Authoritative**: Data that comes from trusted sources or is verified by experts. This can include data such as published books, peer-reviewed research, and in-depth government reports.

- **Unique**: Data that contains unique, diverse knowledge or perspectives. This can include detail on relevant but esoteric scientific subjects, perspectives from different cultures or communities, and rare languages. What is often not useful is data that is not unique or that is repetitive, such as reports that contain very little new information or offer only boiler plate disclosures.
- **Knowledge and/or logic dense**: Data that is in-depth, specific, and includes complex patterns of thought, including applying deductive and inductive logic. This sort of data is particularly helpful for model training. It can include transcripts of expert testimony, reports that consider complex tradeoffs and/or debate sophisticated policies, in-depth reports about communities, and other similar materials.
- **Recent (or still relevant)**: Generally, the value of data will decrease over time, but the rate of decrease depends on the durability of the value. For example, very old news reports on specific local incidents that have no relationship to long-term trends are often not valuable. The same is largely true for books from the 19th century that cover topics that have significantly advanced, such as medical books. However, older material can be valuable if it represents knowledge that is still relevant, such as 19th century math or physics proofs or literature.
- **Long-context**: We generally find that longer and richer documents have more unique and knowledge- and/or logic-dense information. We also find that short documents often do not have dense knowledge and tend to not be as useful, on average, as longer context documents.

For multi-modal data, we find the following qualities of audio, images, and video to be valuable:

- **Audio**: High-quality audio such that voices and/or sounds are clearly identifiable. High diversity of voices in different contexts, such as policy debates, interviews, conversations, storytelling, etc. Having a validated and time-stamped transcription of spoken audio is also valuable.
- **Images**: High volume of quality, diverse imagery, with a focus on non-natural images, as natural images are already abundant. Researchers have a focus on scientific, medical (*e.g.*, x-rays), and business (*e.g.*, charts and data visualizations) images.
- **Video**: High-volume of quality video with diverse subjects. For example, there is ample video with "talking heads", so a priority is on video with diverse subjects. Having video accompanied by text scene descriptions and transcripts is a bonus. These priorities can inform where government spends resources on making data more accessible, discoverable, and useable. We suggest initially spending resources on accessibility. Here, we think making existing digital content discoverable online and taking rich, knowledge-dense content from libraries and archives, would make the most impact.

Governments should facilitate the creation of new open data commons based on national libraries, government research institutions, national archives, museum archives, and other rich content sources. To focus scare resources, governments could bring in the expertise of AI researchers to sample currently inaccessible data across many inaccessible databases and help identify the most valuable sources.

Tools to facilitate open data access

To help achieve some of the objectives set forth above—particularly, making data accessible, secure, and well-governed—Microsoft offers a suite of tools and services designed to facilitate the recommended standards. Leveraging Microsoft's technologies can help governments implement open and controlled access to data, enable fee-based access, employ robust identity and access management (IAM) solutions, optimize websites for LLMs, and achieve seamless data integration and relevancy. The following outlines how Microsoft addresses these aspects through various platforms and services, providing a robust framework for effective data dissemination.

- **Embracing Standardized Ontologies:** Microsoft supports standardized ontologies such as schema.org or NIEM. Azure Digital Twins uses models to comprehensively describe various domains, including manufacturing, building structures, IoT systems, smart cities, energy grids, and web content.
- **Harmonizing and Linking Internal Ontologies and Vocabularies:** Microsoft's Azure platform provides robust data storage solutions that maintain data integrity, accessibility, and quality. Services like Azure Data Factory, Azure Databricks, and Azure Synapse Analytics enhance data quality and accessibility.
- **Gathering Internal and External Written Documentation of Existing Data Products:** Microsoft introduced a tool called "Guidance" to help users control modern language models more effectively and efficiently. This tool enables the interleaving of generation, prompting, and logical control into a single continuous flow that matches how the language model processes the text.
- **Adopting Data Formats Which Allow for Rich Metadata:** Microsoft's Azure platform supports a variety of data formats that include rich metadata. Services like Azure Data Factory, Azure Databricks, and Azure Synapse Analytics enhance data quality and accessibility.
- **Using Open Standards for APIs with the Ability to Link into Knowledge Graphs:** Microsoft promotes the use of open standards for APIs with the capability to link into knowledge graphs.
- **Improving Guidance and Metadata Around Appropriate Data Usage and Licensing:** Microsoft Purview offers a unified data governance solution to manage and govern on-premises, multi-cloud, and software-as-a-service (SaaS) data. It conducts data discovery, data scanning, and access management at scale.

Conclusion

Government data is one of the most valuable assets for AI in the world and reflective of large pools of population and community circumstances. Making it more accessible, discoverable, and usable by AI researchers on an open basis is essential to unlocking the potential not just of the data and AI, but of the profound possibilities for people, industry, and society. We look forward to working with governments, alongside industry and academia, to help advance open data and the benefits it offers.