MASSACHUSETTS TECHNOLOGY COLLABORATIVE

MA AI Hub – Data Commons Collaborative (DCC) Requirements

Date: July 2, 2025

Table of Contents

1.0 Overview

This Massachusetts AI Hub Data Commons aims to empower AI innovation and research across the state by providing an initial capability for centralized, ethical, and secure access to a wide range of data sets. The outcomes will be tailored to AI developers, researchers, startups, and academic institutions building AI capabilities in sectors important to the commonwealth such as transportation, healthcare, climate, robotics, etc. The development of the Data Commons will focus on:

- Building a website that links users to external data providers (like the Australian Research Data Commons capability) and external synthetic data generation capabilities
- Extending the website to expose curated and hosted data sets on a data platform (e.g., AICR)
- Implementing a synthetic data generation capability on the data platform for use case teams
- Establishing and enforcing data security, governance, and ethical use principles
- Implementing a central ledger for tracking Commons Credits to incentivize contributions, promote efficient and equitable used of shared AI resource, and sustain engagement across the MA AI Hub ecosystem through a transparent, trackable credit-based system

2.0 Purpose and Goals

- **Centralize Access** - Provide a curated, categorized directory of high-quality datasets from global repositories, Massachusetts government data sources, and synthetic data tools. Includes both externally linked data sets and direct access to curated data sets hosted on an MA AI Hub platform
- **Data Curation** - Curate a set of high-value datasets (public and MA) for inclusion in the initial hosted website
- **Synthetic Data Generation Capability** - Implement a synthetic data generation capability for safe data experimentation and model development
- **Support AI Innovation** - Enable AI developers to discover, assess, and leverage datasets and simulations to build real-world AI applications
- **Foster Collaboration** - Encourage collaboration between public, private, and academic sectors by making data resources easily accessible

- **Future Expansion -** Provide an extensible foundation for user-generated dataset contributions, collaborative research, and integration with data hosting services
- **Data Security** - Establish a security framework including user access control, audit trails, etc.
- **Ethical Data Use** - Define and enforce governance and ethical use policies for data usage

3.0 Program Phases and Key Milestone Descriptions

Phase	Milestone	Description
1. Initiation & Design	Project Kickoff	Core team resourced, aligned, scope and goals confirmed, project charter approved
	Detailed Architecture & Governance Defined	Detailed technical design and initial governance policies finalized
	Dataset Curation Criteria Approved	Criteria for dataset inclusion reviewed and approved by stakeholders
	Initial Dataset List Confirmed	Validated preliminary list of hosted and external datasets
2. Infrastructure & Setup	Platform Infrastructure Provisioned	Cloud environments, identity management, and foundational services deployed
	Dev/Test Environments Ready	Infrastructure validated for development, testing, and synthetic tooling
3. Development	Data Integration Complete	Data ingestion pipelines and metadata catalog functional for initial datasets
	Synthetic Data Engine Operational	At least one synthetic data tool configured and producing test outputs
	Website Functional	Website supports browsing datasets, basic search, dataset detail pages
	Admin Dashboard Live	Admin can add/edit datasets and manage submissions via UI
4. Testing & Governance	Platform QA Complete	Functional, performance, and security testing completed across all features
	Governance Framework Operational	Role-based access control, data tiers, and usage agreements implemented
5. Rollout & Adoption	User Onboarding & Training Delivered	Documentation finalized; initial user groups trained and onboarded

Phase	Milestone	Description
	Public Launch	MA AI Hub Data Commons goes live for core audience

4.0 Target Users

- AI/ML developers and researchers
- Academic (K-12, higher education) institutions and research labs
- State government departments and civic tech groups
- Startup and enterprise AI teams
- Nonprofits working with Massachusetts open data

5.0 Key Functional Requirements - External Data Sets and Data Capabilities

5.1 Webpage

- Introduction to the Data Commons Collaborative
- Highlighted dataset collections (Public, Massachusetts, Synthetic)
- Quick search and filter capability
- Recent updates or featured datasets

5.2 Dataset Directory

- Searchable and filterable list of datasets
- Filter by:
    - Data Type: Text, Image, Video, Tabular, GIS
    - Source: Public/Global, Massachusetts State, Synthetic Tools
    - Domain: Healthcare, Environment, Transportation, etc.
    - File Format: CSV, GeoJSON, JSON, etc.
    - License: Open, Academic Use, etc.
- Create Dataset Cards with the following information and abilities:
    - Name, Description, Tags
    - Source/Platform (e.g., Hugging Face, MassGIS, CARLA)
    - Download or view link

5.3 Dataset Detail Pages

- Full dataset metadata (size, source, description, license, categories)
- Related datasets
- External documentation links

5.4 Public Data Sets (see Section 7 below for expanded list)

- Example: Hugging Face, World Bank Open Data, Nasa Earth Data, OpenStreetMap, CDC Public Health Data, CERN Open Portal, etc
- Description of each data set
- Download links, license details, and integration notes

5.5 Synthetic Data Tools Section (see below for expanded list)

- Tools like CARLA Simulator, Synthia, NVIDIA Omniverse
- Description of each simulator or synthetic generator
- Links to download/configure or view GitHub
- Use case examples (e.g., autonomous driving models)

5.6 Massachusetts Public Datasets (see below for expanded list)

- Data examples: MassGIS, MassDOT, MA Executive Office of Energy & Environmental Affairs, Health Data from DPH
- Download links, license details, and integration notes

5.7 Community Dataset Submissions

- User submission and data linking suggestion capture
- Form to capture user feedback and intake new data link suggestions

5.8 Admin Dashboard

- Add/edit/remove datasets and links
- Manage user suggestions/submissions
- Link health checker and validation dashboard
- Role-based access controls

5.9 Additional Requirements

- **Accessibility:** Website should meet accessibility standards set by MassTech
- **Performance:** Page load under 2s, search under 2s for 1,000+ datasets
- **Security:** HTTPS, input sanitization, rate limiting for API access
- **Responsiveness:** Fully responsive for mobile, tablet, and desktop
- **Extensibility:** Modular architecture for adding future APIs or data sources

6.0 Key Functional Requirements – Website Functionality for Hosted Data and Access

6.1 Data Hosting and Access

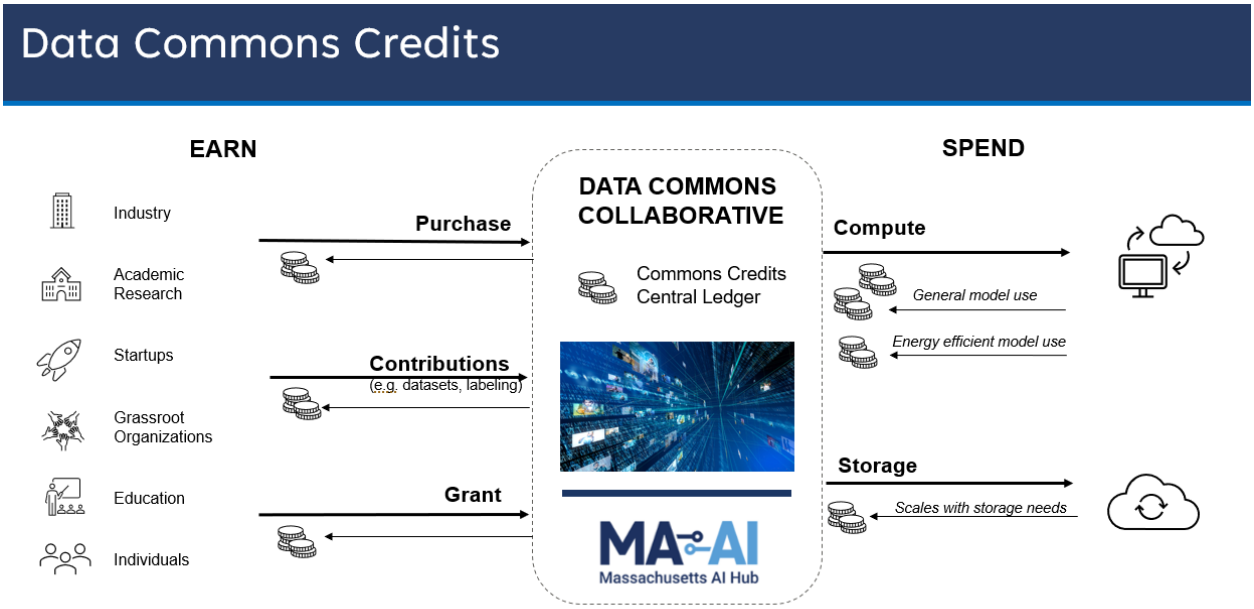
- Implement metadata standards for dataset discovery and context
- Datasets must be searchable, browsable, and downloadable in machine-readable formats (CSV, JSON, Parquet)
- Include APIs and export tools for programmatic data access
- Version control and data lineage tracking must be implemented

6.2 User Experience Enhancements

- Implement a dataset explorer with filters (e.g., category, update frequency, publisher)
- Provide preview functionality (sample rows, schema view).
- Users should be able to request access to restricted datasets via the portal

7.0 Key Functional Requirements – Commons Credits System

Development and implementation of a Commons Credits System to incentivize and manage contributions, usage, and sustainability of the platform. The goal is to create a mechanism that encourages open data sharing, efficient model use, and equitable access across the ecosystem.



7.1 Centralized Ledger and Tracking Infrastructure

- Design and implement a secure, auditable ledger to record the issuance, transfer, and redemption of Commons Credits.
- Ensure compatibility with existing user and contributor identity management systems (e.g., SSO, API tokens, etc.)
- Maintain a historical transaction log with metadata, timestamps, and purpose tags (e.g., “data contribution,” “app access,” “model training”).

7.2 Earning Commons Credits

- Define and implement rules for earning credits through:
  - Contributions: Uploading high-quality datasets, models, tools, or documentation
  - Grants: Receiving credits through the MA AI Hub
  - Purchase: Buying credits with real currency
  - Automate eligibility verification and credit issuance where possible

7.3 Spending Commons Credits

- Enable use of credits for:
  - Accessing premium data sets or models

- Deploying or training models on the Data Commons compute infrastructure
- Purchasing or unlocking third-party applications or services hosted on the platform
- Build in enforcement mechanisms to prevent misuse, including credit expiration, throttling rules, or tiered access.

7.4 Incentivizing Efficient Usage

- Encourage:
  - Low-carbon or energy-efficient model training/inference
  - Use of underutilized infrastructure capacity
  - Sharing of reusable, well-documented assets

7.5 Governance and Transparency

- Provide admin and governance tools for:
  - Reviewing credit issuance policies
  - Setting limits or quotas for users or organizations
  - Generating usage reports and dashboards for stakeholders

7.6 User Experience & Access Control

- Design a front-end portal or integrate with the main Data Commons UI to allow users to:
  - View their credit balance and transaction history
  - Understand how to earn or spend credits
  - Discover resources and apps eligible for credit use
  - Ensure accessibility, multilingual support (where relevant), and support for role-based access (e.g., researcher, developer, public agency staff)

7.7 Technical Expectations

- All components must be modular, secure, and interoperable with the broader Data Commons architecture
- Preference for open-source technologies or frameworks that promote extensibility and transparency
- Compliance with state data security and privacy standards is required
- Support for future blockchain integration or audit trails is a plus, though not a current requirement

8.0 Data Catalog and Metadata Management System

8.1 Capability Definition and Benefits

A data catalog and metadata management system needs to be implemented to manage hosted data that is curated and loaded to the data platform. This section represents the requirements for this capability. The solution shall organize, describe, and make datasets discoverable and reusable by managing their associated metadata. It enables users to publish, search, and access data with clearly defined context—such as who created it, what it contains, how it can be used, and where it came from. This is especially useful in a data commons because it supports FAIR data principles (Findable, Accessible, Interoperable, Reusable), facilitates collaboration and governance, and ensures that data is properly documented, preserved, and cited over time.

The following subsections outline the required capabilities for the metadata management system:

8.2 Metadata Management

- Descriptive Metadata: Title, author, abstract, keywords, and data use/license
- Technical Metadata: File format, schema, size, date of creation/update
- Provenance: Where the data came from, who processed it, and how
- Support for Metadata Standards: Dublin Core, Schema.org, DCAT, etc

8.3 Dataset Discovery and Search

- Faceted Search: Filter by tags, creator, date, format, etc
- Full-Text Search: Search across metadata and even file content
- APIs: Programmatic search and retrieval

8.4 Access Control and Governance

- Role-Based Permissions: Define who can view, edit, publish, or delete data
- Access Tiers: Public, registered, or restricted access levels
- Audit Trails: Logging who accessed or changed what and when

8.5 Dataset Publishing and Versioning

- Data Deposit: Easy upload or ingestion of datasets
- Publishing Workflow: Review and approval process
- Version Control: Track changes across dataset versions

8.6 Interoperability and Integration

- APIs and SDKs: RESTful APIs to integrate with external systems (e.g., analytics tools, cloud storage, Jupyter notebooks, etc.)
- Linked Data / FAIR Compliance: Findable, Accessible, Interoperable, Reusable
- Integration with digital object identifiers (DOIs): Support for persistent identifiers (e.g., DataCite)

8.7 Data Usage and Metrics

- Download and Access Metrics: Who's using the data, how often, etc.
- Citation Tracking: Link data to publications or usage in research

8.8 Collaboration and Community Features

- Annotations and Comments: Allow users to leave feedback or notes
- Team-Based Curation: Support for collaborative dataset management

8.9 Data Previews and Exploration

- In-browser Previews: View CSV, JSON, images, etc. without downloading
- Data Profiling: Summary statistics or sample visualizations

8.10 Example Data Management and Catalog Capabilities for Consideration

- Dataverse, CKAN, DataHub, Amundsen

9.0 Dataset Curation Plan

9.1 Criteria for Inclusion

Datasets included in the inital Data Commons must meet the following criteria:

- High value for AI/ML research
- Extensively used data sets in AI work
- Freely shareable under open or restricted public license
- Non-sensitive or anonymized, or appropriate for synthetic generation

9.2 Suggested Initial Datasets

These suggested initial datasets will need to be vetted by domain experts to ensure they meet the criteria identified above. They are categorized by the following themes:

- General Purpose and Benchmark Datasets

- Natural Language Processing (NLP)
- Health and Life Sciences
- Multimodal and Speech
- Autonomous Systems and Robotics
- Social, Demographic, and Urban Data
- Massachusetts Specific Public Datasets

General-Purpose and Benchmark Datasets

Dataset	Description	Access Type	Approx. Size	URL
ImageNet	Large-scale image classification dataset with over 14 million images across 21,000+ categories.	Licensable	~150 GB	image-net.org
COCO	Object detection and segmentation dataset with 330,000 images and over 1.5 million object instances.	Open	~25 GB	cocodataset.org
MNIST	Handwritten digits dataset containing 70,000 28x28 grayscale images for digit classification.	Open	~50 MB	Kaggle - MNIST
CIFAR-10/100	Image classification datasets with 60,000 32x32 color images in 10 or 100 classes.	Open	~170 MB	CIFAR - Toronto
Open Images	Annotated image dataset with ~9 million images labeled with image-level labels and bounding boxes.	Open	~561 GB	OpenImages GitHub

Natural Language Processing (NLP)

Dataset	Description	Access Type	Approx. Size	URL
Wikipedia Dumps	Raw text from Wikipedia articles, useful for NLP tasks like language modeling and entity recognition.	Open	~24 GB (compressed)	Wikimedia Dumps
Common Crawl	Massive web crawl data including raw web page data, metadata, and text extracts.	Open	Petabytes	Common Crawl
SQuAD	Reading comprehension dataset with over 100,000 question-answer pairs on 500+ articles.	Open	~125 MB	SQuAD Explorer
GLUE	Benchmark suite for evaluating NLP models across nine different tasks.	Open	~33 MB	GLUE Benchmark

The Pile	An 886 GB diverse, open-source dataset of English text created for training large language models.	Open	~886 GB	EleutherAI - The Pile

Health and Life Sciences

Dataset	Description	Access Type	Approx. Size	URL
MIMIC-III/IV	De-identified health-related data from over 40,000 ICU patients, including demographics, vital signs, and more.	Licensable	~100 GB	MIMIC - PhysioNet
PhysioNet	Repository of physiological signals and clinical data, including ECG, EEG, and other recordings.	Open*	Varies	PhysioNet
TCGA	Comprehensive cancer genomics dataset with over 20,000 tumor and normal samples across 33 cancer types.	Open*		NCI - TCGA
UK Biobank	Biomedical database with genetic, lifestyle, and health information from 500,000 UK participants.	Licensable		UK Biobank Access
CheXpert	Large dataset of chest X-rays with uncertainty labels and radiologist-labeled reference standard evaluation sets.	Open*	~500 GB	Stanford CheXpert

*Some datasets require credentialing or data use agreements.

Multimodal and Speech

Dataset	Description	Access Type	Approx. Size	URL
LibriSpeech	Corpus of approximately 1,000 hours of read English speech for automatic speech recognition tasks.	Open	~60 GB	LibriSpeech
AudioSet	Dataset of over 2 million human-labeled 10-second sound clips drawn from YouTube videos.	Licensable	~2 TB	Google AudioSet
YouCook2	Multimodal dataset of cooking videos with temporal annotations for action recognition and captioning.	Open	~1.5 TB	YouCook2

VQA	Visual Question Answering dataset containing images and associated questions to assess visual understanding.	Open	~25 GB	VQA

Autonomous Systems and Robotics

Dataset	Description	Access Type	Approx. Size	URL
KITTI	Benchmark suite for autonomous driving, including stereo, optical flow, visual odometry, and 3D object detection.	Open	~180 GB	KITTI
nuScenes	Large-scale autonomous driving dataset with 3D sensor data and annotations for 1,000 driving scenes.	Licensable	~1.4 TB	nuScenes
CARLA	Open-source simulator for autonomous driving research, providing various urban layouts and environmental conditions.	Open	Varies	CARLA
OpenAI Robotics	Datasets from robotic manipulation tasks, including videos and sensor data for learning-based control.	Open*	Varies	OpenAI Research

Social, Demographic, and Urban Data

Dataset	Description	Access Type	Approx. Size	URL
U.S. Census	Comprehensive demographic data collected by the U.S. Census Bureau, including population, housing, and economic information.	Open	Varies	Census.gov
ACS	American Community Survey providing detailed population and housing information annually.	Open	Varies	ACS
NYC Taxi Data	Ride-level public transportation data, including pickup and drop-off locations, timestamps, and fare amounts.	Open	~100 GB	NYC TLC Trip Data

Massachusetts Specific Public Datasets

Dataset	Description	Access Type	Approx. Size	URL
MA Dept of Public Health Datasets	Health, environment, opioid trends	Open*	Varies by data set	MA Public Health
MBTA Data	Transit schedules and real-time feeds	Open	1 to 10 GB	MBTA
Boston Open Data Portal	City operations, housing, safety, education	Open	10 to 100+ GB	Boston Open Data Portal

* Some may require request or registration

10.0 Synthetic Data Generation Capability

10.1 Capability Definition and Benefits

- Artificially generated information / data that mimics the statistical properties and structure of real-world data without containing any actual personal or sensitive information. Also used to generate additional data where data is sparse (e.g., image data – rotation, color shifting, etc.) Used to enable model training, testing, and validation in privacy-preserving and scalable ways.
- Benefits include enhanced privacy, improved data access where real data is scarce or restricted, and the ability to simulate rare or edge-case scenarios to strengthen model robustness.

10.2 Synthetic Data Generation Requirements

- **Data Ingestion & Preprocessing**
  - Support ingestion of diverse data types (structured, unstructured, time-series, image, etc.) through secure and scalable pipelines
  - Implement automated data profiling and quality assessment tools to evaluate suitability for synthetic data generation
  - Apply de-identification and privacy risk scoring prior to synthetic processing
- **Synthetic Data Generation Capabilities**
  - Provide integrated tools for synthetic data generation using statistical models, generative AI (e.g., GANs, VAEs), and domain-specific simulators

- Allow configurable synthesis parameters (e.g., fidelity vs. privacy tradeoffs, target variable preservation)
- Ensure traceability and metadata linkage between original and synthetic datasets for audit and validation purposes

- **Data Storage and Tagging**
  - Store synthetic datasets with clear metadata tags distinguishing them from real data
  - Record provenance data, synthesis method, and model parameters used for generation
- **Compute Infrastructure**
  - Provision scalable GPU/CPU environments optimized for training generative models and running simulations
  - Enable containerized execution environments for reproducible synthetic data pipelines
- **Privacy and Compliance**
  - Enforce synthetic data generation policies (e.g., HIPAA, etc.) and state data privacy laws
  - Include tools to validate differential privacy levels or other synthetic privacy guarantees
- **Access Control and Governance**
  - Define roles and permissions specific to synthetic data users (e.g., synthetic data creators, consumers, validators)
  - Enable synthetic datasets to be shared more broadly under relaxed governance, while still ensuring auditability

10.3 Synthetic Data Generation Tools Samples

Tool/Framework	Description	Open Source / Licensed

SDV (Synthetic Data Vault)	Deep learning and probabilistic models for tabular data (CTGAN, TVAE, Copula). Developed by MIT	Open Source
MOSTLY AI	High-fidelity synthetic data with built-in privacy and compliance controls; suited for enterprise	Licensed (Commercial)
Gretel.ai	APIs and tools for generating structured/unstructured data with differential privacy	Both (Free tier & Commercial)
YData Synthetic	Toolset for generating tabular and time-series synthetic data with ML and privacy validation	Both (Open Source core, Commercial features)
Synthea	Generates realistic synthetic health records (EHRs) in HL7 FHIR format	Open Source
DataSynthesizer	Lightweight Python library for privacy-preserving synthetic data generation	Open Source (MIT License)
Hazy	AI platform that specializes in enterprise-grade synthetic data for financial services	Licensed (Commercial)
SageMaker Ground Truth Synthetic	AWS-managed service for creating synthetic image and video datasets using 3D simulations	Licensed (AWS Service)

11.0 AI Fairness and Bias Capability

11.1 Purpose

The AI Fairness and Bias Capability is intended to ensure that data and AI systems developed within the Data Commons are equitable, transparent, and aligned with ethical standards. It supports identifying, mitigating, and monitoring bias in datasets, models, and outcomes, particularly in public sector and high-impact applications.

Human-Centric

AI and data governance should prioritize the rights, dignity, and well-being of individuals. Systems must be designed to support and enhance human autonomy and avoid replacing or undermining human agency. The DCC must ensure that technologies developed and deployed serve the public good and contribute to societal benefit.

Trustworthy

The DCC must foster confidence in AI systems and data management through rigorous attention to security, reliability, explainability, and transparency. Mechanisms must be in

place to ensure accountability, uphold ethical norms, and maintain compliance with legal standards, so users and the public can trust that systems will behave as expected.

Inclusive

Governance of the DCC must actively address and reduce inequities by ensuring that datasets, models, and applications reflect a diverse range of perspectives and experiences. It must mitigate bias, prevent discrimination, and promote equitable access to participation and benefit from AI-driven insights and innovations.

Innovative

The DCC must enable cutting-edge research and development by promoting open collaboration, cross-sector partnerships, and access to high-quality data resources. By creating a platform that supports experimentation and rapid iteration, it will help position Massachusetts as a global leader in ethical AI innovation.

Sustainable

Data and AI infrastructure must be designed and managed with attention to environmental impact and long-term viability. The DCC should support practices that conserve resources, promote energy efficiency, and respect planetary boundaries while considering intergenerational justice and long-term societal benefit.

12.0 Governance Requirements

Legal and Regulatory Guidance

The DCC must adhere to state and federal laws regarding data use, access, and sharing. Clarity regarding the legal requirements and obligations are required. Some potential requirements include:

- Ensuring fair use compliance and avoiding unauthorized commercialization
- Observing anti-kickback statutes to prevent unethical financial incentives in data sharing .
- Implementing nondiscrimination policies that protect individuals from disparate treatment in AI outcomes
- Structuring data-sharing agreements under a clear "Data as a Service" model with delineated provider-user responsibilities

Ethical Considerations

Ethics must be embedded across the DCC lifecycle:

- Develop standardized methods for quantifying and reducing algorithmic bias
- Conduct regular fairness assessments, particularly in high-risk or sensitive domains
- Implement privacy risk mitigation tools, such as differential privacy, k-anonymity, and secure multi-party computation
- Design systems that provide transparency in AI decision-making and access to legal remedies for affected individuals
- Guard against the risk of identification when external data is combined with DCC datasets by managing quasi-identifiers (see platform requirements document)

Data Sharing Models

DCC will support multiple levels of access based on sensitivity and user roles:

- Federated Modeling and Validation: External entities send models to DCC for training/testing without accessing raw data
- Restricted Access: Highly secure systems that allow in-person analysis with audit logs
- Controlled Access: Remote, role-based access enabled through an authentication and approval process
- Federated Access: Enables model execution on datasets stored remotely with defined visibility
- Open Access – Need to determine if this capability will be permitted (e.g., non-sensitive data is publicly available for download and exploration without barriers

Consent Management

Consent mechanisms must be dynamic and verifiable:

- Informed consent processes must explain data uses, rights, and opt-out options
- All consent must be tokenized, revocable, and time-bound to ensure clarity and traceability
- Systems must purge access rights and data derivatives upon revocation or expiration of consent

Governance Model

The DCC shall employ a hybrid governance model, including:

- Centralized coordination via a Data Access & Use Committee for shared or publicly held data.
- Distributed approval frameworks for data held by partner organizations, with brokering functions.
- Licensing regimes tailored to dataset modality and sensitivity (e.g., biometric, PHI).
- Transparent documentation of approval processes, opt-out policies, and data stewardship practices.

Example Roles and Access Controls

Role	Description
Administrators	Oversee system management, role access assignment, and overall governance enforcement.
Data Developers	Curate, standardize, redact, and prepare data for use within the DCC. Responsible for quality assurance.
Data Users	Researchers, analysts, or institutions granted access to specific datasets under policy and consent guidelines.
Challenge Participants	Time-limited access for competitions, sprints, or collaborative innovation labs.
Auditors	Independently assess system compliance, privacy, and ethical safeguards.
Legal Cousel	Provides direction according to regulations and laws
Policy Analysts	Reviews and aligns with emerging regulations

12.1 Core Capabilities

Dataset Auditing and Bias Detection

- The system should allow automated and manual audits of datasets to detect potential biases across protected attributes such as race, gender, age, and

disability. It should support statistical fairness metrics and provide visualizations to highlight disparities and areas of concern. Datasets lacking sufficient representation or posing risks for reinforcing inequities should be flagged for review.

Model Fairness Evaluation

- Fairness evaluation tools should be integrated into the AI/ML model development workflow. Users must be able to assess and compare model performance across different subpopulations and document the outcomes.

Bias Mitigation Tools

- The platform should offer a comprehensive set of mitigation tools, including pre-processing (e.g., reweighting), in-processing (e.g., fairness-aware learning algorithms), and post-processing (e.g., outcome adjustment techniques). Users should be able to explore and document the trade-offs between fairness and performance when applying these techniques.

Ethical Oversight and Review

- A formal workflow should support ethical review of datasets and models, including annotations and approvals by reviewers. Projects using Commons data should include bias impact statements and ethical considerations as part of their documentation and review process.

Community and Stakeholder Engagement

- To ensure equitable outcomes, the platform should include mechanisms for involving impacted communities in the audit and review process. This could include participatory audits, community feedback channels, and documentation of social and historical context for data and models.

12.2 Operational Requirements

Transparency and Auditability

- All fairness assessments, mitigation efforts, and decisions must be logged and made auditable. Public summaries of fairness evaluations should be generated to foster transparency and accountability in the use of the Data Commons

Usability

- Fairness and bias tools should be accessible and user-friendly, with intuitive interfaces and comprehensive documentation. The system should provide

recommended workflows and default settings for users who are not fairness experts.

Privacy and Security

- Bias audits should respect data privacy, ensuring sensitive attributes are protected and access-controlled. Fairness evaluations should not expose individual-level data or increase re-identification risk.

12.3 Governance and Legal Compliance

- The capability should align with state and federal anti-discrimination laws, as well as emerging national and international AI regulations such as the Algorithmic Accountability Act and the EU AI Act. Governance structures should define roles for fairness auditors, data stewards, and reviewers, with a designated Fairness Oversight Board for high-risk applications. (see Governance requirements doc)

12.4 Technical Integration

- The tools should integrate with popular AI development environments and workflows (e.g., Jupyter, MLFlow, etc.) APIs should allow programmatic access to fairness functions. Integration with open-source fairness libraries (e.g., Fairlearn and AIF360 – see below or other tool considerations) is needed, as is support for ongoing monitoring of deployed models (within the rapid prototyping environment – read "not a production inference / model scoring environment") for fairness drift.

12.5 Monitoring and Reporting

- The system should automatically generate fairness reports for both datasets and models. Dashboards should be provided to monitor fairness metrics across projects, with a version history of audits and mitigation actions. Alerts should be triggered if models show signs of performance or fairness drift over time.

12.6 Education and Capacity Building

- Training and education modules should be available to help users understand fairness concepts, tools, and ethical considerations. The platform should host

regular workshops and maintain a repository of case studies and best practices to build a community of responsible AI practitioners.

12.7 Potential AI Biase and Fairness Open Source Tools for Consideration

Capability Area	Tool	Description
Dataset Auditing	AIF360	Dataset bias detection, fairness metrics, and mitigation algorithms
	Fairlearn	Fairness assessment and mitigation with visualizations and dashboards
	Themis-ML	Measures group and individual fairness in models
	DataPrep.EDA / Pandas Profiling	Automated profiling of datasets for distributional skews, outliers, and missing values
Bias Mitigation	AIF360	Includes pre-, in-, and post-processing mitigation algorithms
	Fairlearn	Focuses on in-processing (reductions) and post-processing constraints
	Adversarial Debiasing	Implemented in TensorFlow/PyTorch; trains models to be both accurate and fair
	Jurity	Provides implementations of fairness metrics (e.g., Demographic Parity, Equal Opportunity, Predictive Parity, etc.)
Model Explainability	What-If Tool (WIT)	Interactive tool for exploring model predictions and fairness scenarios
	SHAP	Model-agnostic tool for feature importance (local and global)
	LIME	Explains individual model predictions locally.
Dataset Documentation	Datasheets for Datasets	Templates for documenting provenance, ethics, and context of datasets
	Open Ethos	Tools for documenting ethical risk and engaging stakeholders
Monitoring & Drift Detection	Evidently AI	Monitors data quality, bias, and model performance over time
	WhyLogs	Logs statistical metrics for real-time drift and fairness monitoring

Capability Area	Tool	Description
Integration / Pipelines	MLFlow + Fairlearn	Tracks models and supports fairness evaluations in development pipelines
	TensorBoard + WIT	Combines model training visualization with fairness evaluation
	Jupyter Notebooks	Common interface for running most fairness tools interactively
	Airflow	Orchestrates fairness monitoring and audit pipelines

13.0 Security and Access Control Requirements

13.1 Access Controls

- **User Authentication**:
  - Single sign-on via identity capability
  - Multi-factor authentication for sensitive data access
- **Role-Based Access Control (RBAC)**:
  - Roles: Public Viewer, Researcher, Curator, Admin
  - Access to restricted datasets granted only via approval workflow
- **Dataset Permissions**:
  - Public, Restricted (requires review), Internal (e.g., MA agencies)
- **Logging & Auditing**:
  - Log all data access, downloads, synthetic generation events
  - Retain logs for 3+ years for compliance review

13.2 Data Protection

- Encrypt data at rest and in transit
- Integrate intrusion detection and monitoring
- Perform regular security reviews and penetration testing

14.0 Governance Policies and Ethical Guidelines

14.1 Guiding Principles

- Transparency: Public documentation on data sources, transformations, and synthetic generation methods
- Accountability: Track data access and enforce responsible usage
- Equity: Avoid inclusion of biased or non-representative data
- Consent and Compliance: Datasets must comply with applicable laws

Appendix A - Expanded List of External Links for Data Sets and Synthetic Tools

A.1 Public Data Sets

Name	Description	Domain	Url
Amazon Web Services (AWS) Public Datasets	Datasets hosted on AWS for various research and analysis purposes.	Misc	Open Data on AWS
AWS Open Data Registry	Cloud-hosted large-scale datasets	AI, Big Data, Climate	AWS Open Data
Bitcoin Open Dataset	Blockchain transactions and network activity	Blockchain	Blockchain Data
Cancer Research Data Commons	Datasets from National Cancer Institute and NIH funded programs as well as key external cancer programs	Healthcare	Datasets | CRDC
CDC Public Health Data	Health and disease-related datasets	Healthcare	CDC Data
CERN Open Data Portal	Particle physics datasets from the Large Hadron Collider	Physics	CERN Open Data
Cityscapes Dataset	Urban street scene dataset for segmentation tasks	AI, Computer Vision, Urban	Cityscapes
CVE (Common Vulnerabilities and Exposures)	Publicly disclosed cybersecurity vulnerabilities	Cybersecurity	CVE
Data for the Common Good	Pediatric Cancer	Healthcare	Data for the Common Good
Data.gov	U.S. government open data repository	Government	Data.gov
Data.gov.uk	UK government's open data portal with datasets on demographics, government spending, and more.	Government	Data.gov.uk
Data.world	Social network for data people to discover and share datasets.		Data.world
DataHub	Platform for discovering and sharing datasets.		DataHub
Enigma Public Data	Collection of public datasets for various industries and topics.		Enigma Public Data
Enron Email Dataset	Corporate email dataset for NLP research	Cybersecurity, NLP	Enron Data

Name	Description	Domain	Url
ETH Blockchain Dataset	Transaction history for Ethereum blockchain analysis	Blockchain	Etherscan
FAO Data Portal	Agriculture, food security, and nutrition data	Agriculture, Food	FAO Data
FBI Crime Data	U.S. crime statistics and reports	Law, Crime	FBI Crime Data
FiveThirtyEight Data	Journalism-driven analytics and datasets	Politics, Society, Economics	FiveThirtyEight
FiveThirtyEight Data	Datasets used in FiveThirtyEight's data journalism and analysis.		FiveThirtyEight Data
Genomic Data Commons	National Cancer Institute repository of cancer genomic studies	Genomics	https://gdc.cancer.gov/about-gdc
GitHub	Public datasets on programming languages, software development trends, and more.		GitHub
GitHub Awesome Public Datasets	Curated list of high-quality public datasets.		Awesome Public Datasets
Global Health Observatory Data Repository	World Health Organization's data on global health indicators.		WHO Data
Google Data Commons	Public data from a wide variety of sources covering topics including Agriculture, Biomedical, Crime, Demographics, Economy, Education, Energy, Environment, Health, and Housing	General	Home - Data Commons
Google Dataset Search	Search engine for publicly available datasets	General	Google Dataset Search
Harvard Dataverse	Repository for sharing, citing, and exploring research data.		Harvard Dataverse
HealthData.gov	U.S. health-related datasets	Healthcare	HealthData
HotpotQA	Question-answering dataset for NLP	AI, NLP	HotpotQA
Hugging Face Datasets	NLP and AI model training datasets	AI, NLP	Hugging Face
IMF Data	Economic and financial datasets	Finance, Economy	IMF Data

Name	Description	Domain	Url
Inside Airbnb	Data on Airbnb listings, pricing, and availability	Real Estate, Tourism	Inside Airbnb
Ionic Liquids Database (ILThermo)	Thermodynamic properties of ionic liquids	Chemistry	ILThermo
Kaggle Datasets	Open ML repository with datasets across domains	General, AI, Finance, Health	Kaggle
Massachusetts Data Hub	Data and reports published by Mass state agencies	Government	https://data.mass.gov/
NASA Earth Data	Open satellite imagery and climate datasets	Space, Climate	NASA
NOAA Climate Data	Historical and real-time weather datasets	Climate, Weather	NOAA
NOAA Data	Environmental and weather datasets for research and analysis.	Enviro	NOAA Data
OECD Data	International development, economic trends, and social issues.		OECD Data
Open Images Dataset	Large-scale image dataset for computer vision research	AI, Computer Vision	Google Open Images
OpenML	Platform for sharing datasets and machine learning experiments.		OpenML
OpenStreetMap	Global geographic mapping data	GIS, Transportation	OpenStreetMap
Penn State Data Commons	Repository of Penn State research data across all subjects	Government	Penn State Data Commons
Psychiatric Genomics Consortium (PGC)	Data on mental health and genetics	Neuroscience, Genomics	PGC Data
PubChem	Open database of chemical molecules and properties	Chemistry	PubChem
Quandl	Financial, economic, and alternative datasets	Finance, Economy	Quandl
Sanger COSMIC Database	Catalog of somatic mutations in cancer	Genomics, Cancer Research	Sanger COSMIC
Sequence Read Archive (SRA)	Raw sequence data for genomics research	Genomics	SRA Data
Stanford Large Network Dataset	Social and information networks	Social Science, AI	Stanford SNAP
The Cancer Genome Atlas (TCGA)	Genomic data from cancer research	Genomics, Cancer Research	TCGA

Name	Description	Domain	Url
Therapueutics Data Commons	AI ready datasets for drug discovery	Healthcare	https://tdcommons.ai/
UCI Machine Learning Repository	Curated ML benchmark datasets Repository of datasets commonly used in machine learning research and analysis.	AI, ML Research	UCI ML Repository
UK Open Data (data.gov.uk)	UK government open datasets	Government	UK Data
UN Data Portal	International statistics from UN agencies	Global Development, Social	UN Data
UN Office on Drugs and Crime	Global crime and corruption statistics	Law, Crime	UNODC
Web Data Commons	Web Data Commons project extracts structured data from the Common Crawl, the largest web corpus available to the public, and provides the extracted data for public download in order to support researchers and companies in exploiting the wealth of information that is available on the Web.	General	Web Data Commons
World Bank Open Data	Economic, development, and global financial data	Finance, Economy	World Bank
Yelp Open Dataset	Business reviews and check-in data	Business, Marketing	Yelp Dataset

A.2 Massachusetts Public Data Sets

Name	Description	Url
Massachusetts Open Data Portal	Official state government data covering various categories such as economy, health, transportation, and public safety.	data.mass.gov
Massachusetts GIS (MassGIS)	Geographic and spatial data including land use, transportation, and environmental features.	mass.gov/orgs/massgis-bureau-of-geographic-information
Boston Open Data	Open datasets from the City of Boston, including public safety, infrastructure, education, and permits.	data.boston.gov

Name	Description	Url
Cambridge Open Data Portal	City-level data for Cambridge, covering housing, business, and city services.	data.cambridgema.gov
MBTA (Massachusetts Bay Transportation Authority) Data	Real-time and historical public transit data, including bus and train schedules.	mbta.com/developers
Massachusetts Public Records Request Database	Information on state public records requests and open government.	mass.gov/massachusetts-public-records-law
Massachusetts Department of Public Health (MDPH) Data	Health statistics, COVID-19 reports, birth and death records, and disease surveillance.	mass.gov/orgs/department-of-public-health
Massachusetts Environmental Public Health Tracking	Environmental health data, including air quality, cancer rates, and disease tracking.	matracking.ehs.state.ma.us
Community Health Information Profile (MassCHIP)	Data on health indicators at the community level.	mass.gov/service-details/masschip-data-available
Massachusetts Department of Elementary and Secondary Education (DESE) Data	School performance, enrollment statistics, and standardized test scores.	profiles.doe.mass.edu
Massachusetts Higher Education Data Center	Data on higher education institutions, graduation rates, and tuition costs.	mass.edu/datacenter/home.asp
Massachusetts Labor Market Information (LMI)	Employment trends, wage statistics, and labor force participation data.	lmi.dua.eol.mass.gov/LMI/Home
Massachusetts Department of Environmental Protection (MassDEP) Data	Air and water quality data, hazardous waste sites, and environmental permits.	mass.gov/orgs/massachusetts-department-of-environmental-protection
Massachusetts Energy and Environmental Affairs (EEA) Data Portal	Renewable energy, conservation, and climate change impact data.	mass.gov/orgs/executive-office-of-energy-and-environmental-affairs
USGS Massachusetts Water Science Center	Water quality, streamflow, and groundwater data for Massachusetts.	usgs.gov/centers/ma-water
Massachusetts Crime Statistics (EOPSS)	State-level crime reports, public safety data, and law enforcement statistics.	mass.gov/orgs/executive-office-of-public-safety-and-security

Name	Description	Url
Boston Police Department Crime Data	Crime incidents, locations, and historical trends.	data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system
Massachusetts Housing Data Portal	Affordable housing statistics, mortgage programs, and rental market trends.	masshousing.com
Boston Housing Authority Data	Public housing data and low-income housing assistance information.	bostonhousing.org
Zoning & Land Use Data (MassGIS)	Land use planning, zoning maps, and urban development data.	mass.gov/info-details/massgis-data-land-use-2016

A.3 Synthetic Data Tools

Dataset Name	Description	URL
SYNTHIA by MITRE	Synthetic dataset for autonomous vehicle training with street scenes and driving scenarios.	https://synthea.mitre.org/
CARLA Simulator	Open-source simulator for autonomous driving systems with RGB images, LiDAR, and depth maps.	CARLA
Tonic	Platform for generating synthetic data based on real datasets	Tonic
Senthetic Data Vault	Synthetic tabular data and relational datasets using statistical methods to preserve data privacy.	SDV.DE
Unity Perception	Toolkit for generating synthetic data for computer vision tasks, including object detection and segmentation.	Unity Perception
SceneNet RGB-D	Synthetic dataset providing RGB-D images with labeled objects for indoor scene understanding.	SceneNet RGB-D
AirSim by Microsoft	Open-source simulator for drones and autonomous systems, providing high-fidelity synthetic data.	AirSim
DeepDrive	Platform for training autonomous systems with synthetic data and simulation tools.	DeepDrive
GTAV Dataset	Synthetic dataset generated from the *Grand Theft Auto V* game for tasks like object detection and scene understanding.	GTAV Dataset

Microsoft COCO (Synthetic)	Synthetic extension of the COCO dataset for real-world image-like data generation for AI training.	Microsoft COCO
Sim4CV	Synthetic data generator for computer vision tasks, offering labeled images and videos.	Sim4CV
UnrealCV	Plugin for Unreal Engine to generate synthetic data for vision-based applications.	UnrealCV
Virtual KITTI	Synthetic dataset created from *KITTI* benchmark suite simulations for autonomous driving.	Virtual KITTI
IBM Synthetic Data Vault	Framework for generating synthetic tabular data mimicking real-world datasets while preserving privacy.	IBM SDV
DataGen	Platform for generating synthetic 3D data for computer vision tasks like object detection and segmentation.	DataGen
Google Research – Generative Models	Google's generative models, including GANs, for creating synthetic data like images and videos.	Google Research GANs
OpenAI's Dactyl	Synthetic dataset for training robotic hands using reinforcement learning in simulated environments.	OpenAI Dactyl
Synthetic Financial Datasets	Publicly available synthetic financial data for modeling, fraud detection, and market predictions.	Synthetic Financial Data