Data Commons Collaborative RFP Questions and Answers

Contents

Administrative/Procurement

Question 1: Section 3.1(d) requires confidential material be included only in hard copy, yet 3.1(a)/(c) requires electronic submission to proposals@masstech.org. How should respondents submit confidential content given electronic-only submission?

Answer: Please note that no information deemed confidential by an applicant should be submitted in any form unless the MassTech General Counsel has approved a request for confidential treatment. If an applicant's request for confidential information has been approved, the applicant should submit the confidential information in hard copy only to the Procurement Team Leader at the address on the RFP cover page.

Question 2: Payment Terms: What are the standard payment terms (e.g., net 30) and preferred invoicing schedule (e.g., monthly upon achievement of milestones)?

Answer: Payment terms are typically on a cost reimbursement basis. Participants shall submit invoices to MassTech no more frequently than monthly and no less frequently than quarterly. Payments will be made by MassTech within 30 days following receipt of an invoice submitted in a format consistent with the contract requirements.

Question 3: Is there a preferred format for the proposal submission (e.g., memo vs. PowerPoint)? Will there be an opportunity to present the proposal materials live?

All financial submissions must use the provided budget template. Proposals may be submitted in Word and/or PowerPoint format. If MassTech requests oral presentations, vendors should be prepared to present using PowerPoint and/or provide a demonstration.

Question 4: If we intend to submit confidential materials, can you confirm whether a hard copy submission is required in addition to the electronic submission?

Please note that no information deemed confidential by an applicant should be submitted in any form unless the MassTech General Counsel has approved a request for confidential treatment. If an applicant's request for confidential information has been approved, the applicant should submit the confidential information in hard copy only to the Procurement Team Leader at the address on the RFP cover page.

AI & Analytics Capabilities

Question 5: LLM & API Integration - As the platform is described as 'extensible' and designed to support innovation, do you expect it to connect to popular AI tools like OpenAI or open-source models (for example, using APIs from Cohere, Mistral, or Hugging Face Transformers)?

Answer: Yes, the system should have the ability to connect to data, capabilities, APIs outside the system

Question 6: Infrastructure Management - Because the platform must support 'logging, encryption, and audit trails,' will infrastructure observability and lifecycle management be in scope—like centralized monitoring, patching, and compliance reporting (for example, using infrastructure automation tools or cloud-native monitoring platforms)?

Answer: Yes. To meet the requirements for logging, encryption, and audit trails, infrastructure observability and lifecycle management will be in scope. The platform will need to have implemented a centralized monitoring and alerting (covering APIs, data services, and network flows), alongside patch and configuration management to ensure system security and reliability. Compliance reporting will be supported through integrated logging pipelines, audit dashboards, and regular policy checks tied to access controls and dataset tiering. These functions will be delivered using infrastructure automation and/or cloud-native monitoring tools to provide consistency across environments.

Question 7: Should the platform support fairness checks for models in addition to datasets?

Answer: Yes — the platform should support fairness checks for models as well as datasets.

Question 8: Does the governance platform need to include workflow capabilities?

Answer: Yes, workflow capabilities are valuable for ensuring transparent and auditable governance processes, such as dataset approvals, access requests, and compliance checks.

Question 9: Do you have any Single Sign On (SSO) requirements that would grant users access to other MassTech sites / platforms?

Answer: No, Access to other MassTech sites and platforms is out of scope

Question 10: Which access paths are required for hosted data (bulk files vs. filtered REST/SQL-like)?

Answer: Specific access path requirements have not yet been finalized. The system should be designed to support multiple modes of access—including bulk file transfers as well as filtered query-based methods (e.g., REST APIs or SQL-like interfaces)—with vendors encouraged to recommend approaches that maximize usability, scalability, and security.

Question 11: Have you defined a storage footprint (or similar limits) regarding how much data storage, and supporting cost, you are comfortable hosting?

Answer: A specific storage footprint or hosting limit has not yet been defined. Vendors should propose scalable storage approaches and cost models that accommodate growth while ensuring efficiency and alignment with the goals of the Data Commons.

Question 12: How many environments do you plan to have across the key platforms (dev/test/UAT/ prod)?

Answer: Dev/Test, QA, and Production environments are required

Question 13: Will the DCC maintain any exclusivity over datasets that are submitted by users or can they be hosted in other platforms as well?

Answer: The DCC will not maintain exclusivity over user-submitted datasets.

Question 14: Are there specific fairness/bias frameworks (e.g., AIF360, Fairlearn) that MassTech prefers, or should we propose a toolkit?

Answer: No preferred frameworks exist. Project scope includes recommending, designing, and implementing fairness/bias detection capabilities. Anticipated in this scope are frameworks for data set auditing, bias mitigation, model explainability, drift detection, etc. Frameworks such as AIF360, Fairlearn and others can be included (e.g., Jurity, SHAP, etc.)

Question 15: How should multi-tiered dataset access (Public, Restricted, Internal) be implemented—does MassTech have a reference access model?

Answer: MassTech has not established a reference access model for multi-tiered dataset access at this stage. Vendors should propose an approach to implementing differentiated access levels (e.g., Public, Restricted, Internal) that incorporates strong identity and access management, governance controls, and auditing capabilities, while ensuring flexibility to adapt as policies and requirements evolve.

Question 16: Should metadata management support federated discovery of external datasets (without hosting them) or only curated hosted datasets?

Answer: Curated, hosted datasets within the DCC, and federated discovery of external datasets (without hosting them), provided the metadata catalog makes them findable, includes provenance, and links to the authoritative source.

Question 17: Should logs be integrated with SIEM tools (Splunk, ELK, Azure Sentinel)?

Answer: The system must be designed for strong security and continuous monitoring. Vendors should recommend appropriate approaches—such as integration with SIEM tools (e.g., Splunk, ELK, Azure Sentinel) or equivalent solutions—to ensure robust threat detection, auditing, and incident response.

Question 18: Will the platform require FedRAMP High / NIST 800-53 compliance, or only Massachusetts state cybersecurity framework?

Answer: The required level of security compliance—whether Massachusetts state cybersecurity standards, FedRAMP High, NIST 800-53, or other frameworks—has not yet been determined. Vendors should propose security architectures and controls that meet generally accepted best practices and can adapt to whichever specific compliance standards are finalized during the project.

Question 19: Could you elaborate on your expectations regarding Linked Data and FAIR compliance (Findable, Accessible, Interoperable, Reusable)? Specifically, are there particular standards, practices, or frameworks you would like us to align with?

Answer: Specific standards and practices for Linked Data and FAIR compliance have not yet been defined. Vendors should propose approaches that align with recognized best practices and widely used frameworks—such as FAIR principles, W3C standards, and linked-data vocabularies (e.g., DCAT, Schema.org)—to ensure that datasets are Findable, Accessible, Interoperable, and Reusable as the platform evolves.

Question 20: Is it preferred that the datasets are hosted by the DCC where possible or defer to external links where possible? The implications of which are more control of format/delivery options vs less storage costs

Answer: A balanced approach is desired. High-value, frequently accessed, or shared research/teaching datasets are strong candidates for DCC hosting. Sensitive, very large, or rapidly changing datasets may remain in external repositories, with DCC providing metadata, access governance, and usage tracking through federated discovery.

Question 21: Does MassTech expect the DCC to be built on or integrated with any existing state-managed cloud infrastructure (e.g., MassGIS, MA HIE, Mass Open Data Portal)?

Answer: It is anticipated that the system will integrate with other state AI capabilities including the AI Compute Resource (AICR) being implemented by MGHPCC and Snowflake (MA state solution).

Question 22: Will the selected vendor be expected to design for healthcare interoperability requirements with other Massachusetts or regional data infrastructure initiatives (e.g., Mass Open Cloud, state health systems, research data commons)?

Answer: Yes. The selected vendor should design the solution to accommodate healthcare interoperability requirements and support integration with relevant Massachusetts and regional data infrastructure initiatives (e.g., Mass Open Cloud, state health systems, research data commons), in alignment with applicable privacy and security standards.

Question 23: Are vendors expected to provide or configure hosting environments (e.g., cloud setup) or will MassTech procure infrastructure independently?

Answer: Vendors are expected to propose and, if applicable, configure hosting environment options (e.g., cloud setup) as part of their solution. MassTech will be responsible for procuring the selected infrastructure independently, based on the recommended approach.

Architecture & Infrastructure

Question 24: Hosting Environment - Given that the platform must 'host curated datasets' and support 'synthetic data generation,' will the solution be deployed in a specific environment—like on-premises, in the cloud, or in a hybrid setup (for example, using scalable storage systems, virtualized infrastructure, or container orchestration platforms)?

Answer: The ability to host datasets on a hybrid setup is required.

Question 25: Are there required delivery models like on-prem private cloud, public cloud or hybrid?

Answer: Yes, hybrid. The architecture should be designed to interoperate across these domains with consistent identity, RBAC, logging, and audit. At launch, a hybrid model leveraging MGHPCC/AICR.

Question 26: How many separate environments will need to be provisioned? Examples would be Dev/Test, QA and Production.

Answer: Dev/Test, QA, and Production environments are required

Question 27: Will the website be hosted on a specific platform (e.g., AWS, Azure, MassTech infrastructure)?

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 28: If MassTech requires a technical solution architecture at RFP stage, please provide guidance about how to present this solution and the associated ongoing monthly costs to operate the underlying cloud-based services.

Answer: At the RFP stage, vendors should include a high-level technical solution architecture that demonstrates how the proposed system will meet the functional and non-functional requirements of the Data Commons. The architecture should illustrate major components (e.g., data ingestion, storage, governance, access, security, analytics), their interactions, and reliance on MGHPCC AICR and cloud-based services or third-party tools.

To support evaluation, vendors should provide estimated ongoing costs for operating the proposed solution. These should be presented as line items in the budget template (e.g., cloud compute, storage, networking, licensing, monitoring/observability), with assumptions clearly stated (such as expected data volumes, number of active users, or compute hours). Costs can be shown as monthly or annual recurring expenses, with scalability considerations noted (e.g., how costs change as data or usage grows).

MassTech understands that detailed architecture and pricing will be refined during implementation. At this stage, we are looking for reasonable, transparent estimates that demonstrate vendors’ ability to anticipate and plan for sustainable operations. Vendors may include both baseline costs and optional enhancements or tiers, provided the assumptions are clearly explained.

Question 29: Are there preferred cloud environments or vendors (AWS, Azure, GCP, state-managed infrastructure) for deployment?

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 30: Will the DCC platform be hosted on MassTech infrastructure, or should respondents propose cloud hosting (e.g., AWS, Azure)? Please elaborate on current state of cloud platforms that we would be expected to utilize if hosted within current MassTech environments.

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 31: Does MassTech have a preferred cloud provider (AWS, Azure, GCP) or should vendors recommend?

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 32: Will hosting infrastructure be provided by MassTech, or should the vendor provision/manage cloud infra?

Answer: Yes, hybrid. The architecture should be designed to interoperate across these domains with consistent identity, RBAC, logging, and audit. At launch, a hybrid model leveraging MGHPCC/AICR for hosted data and compute, paired with cloud-based services for catalog, APIs, and web access. The platform should be extensible to multi-cloud or federated deployments as the platform scales and partner institutions contribute additional resources.

Question 33: Should CPU/GPU-based deep generative models (GANs, VAEs) be hosted centrally, or is it acceptable for users to bring their own compute?

Answer: Yes, hybrid. The architecture should be designed to interoperate across domains. At launch, a hybrid model leveraging MGHPCC/AICR for hosted data and compute, paired with cloud-based services for catalog, APIs, and web access. The platform should be extensible to multi-cloud or federated deployments as the platform scales and partner institutions contribute additional resources.

Question 34: Hosting & Cloud Strategy: What is the preferred hosting environment for the DCC? (e.g., MassTech-owned infrastructure, a specific cloud provider like AWS/Azure/GCP, or a hybrid model?). Are there existing state cloud contracts or preferences we should be aware of?

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 35: Can you confirm if there is a preferred hosting environment (state-run vs. AWS/Azure/GCP)

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 36: Are you seeking a customer built dataset explorer or do you have a preferred tool?

Answer: The RFP does not prescribe a specific dataset explorer tool. Instead, it requires a dataset directory with browsing, metadata search, provenance, and detail pages. This can be delivered either through a custom-built interface or by extending established open-source platforms such as CKAN or Dataverse. Our recommendation is to leverage an open-source base for speed and standards compliance, while adding custom modules for features unique to the MA AI Hub—such as Commons Credits, fairness dashboards, and synthetic data integration.

Question 37: What technology stack is preferred or mandated for the website, if any?

Answer: There are no existing preferences or mandated requirements for the technology stack to be used for the website. Vendors are expected to propose a recommended stack based on best practices, scalability, security, and alignment with the goals of the Data Commons platform.

Question 38: Who will be responsible for the ongoing administration of the website after launch?

Answer: Initially Massachusetts AI Hub will be responsible for ongoing admin.

Question 39: The RFP/requirements list examples (CKAN, Dataverse, DataHub, Amundsen). Are these preferences or references? Any constraints (e.g., mandated open-source, licensing) or prior Commonwealth standards we should align to?

Answer: The tools listed in the RFP/requirements (e.g., CKAN, Dataverse, DataHub, Amundsen) are provided as illustrative references and are not preferences or mandated solutions. There are no constraints such as required open-source licensing or existing Commonwealth standards at this time. Vendors are encouraged to recommend platforms and approaches that best align with the goals of the Data Commons, balancing scalability, usability, security, and cost-effectiveness.

Question 40: Are there preferred or mandatory metadata/catalog technologies (e.g., CKAN, Dataverse, Amundsen) that must be used?

Answer: No, there are no preferred or mandatory metadata solutions. Vendors should recommend appropriate standards (e.g., DCAT, Schema.org, or others) and propose a practical minimum completeness bar that balances usability, interoperability, and ease of implementation for the initial launch.

Question 41: Which metadata standards are preferred (Dublin Core, DCAT, Schema.org)?

Answer: No specific metadata standards are mandated at launch. Vendors should recommend appropriate standards (e.g., DCAT, Schema.org, or others) and propose a practical minimum completeness bar that balances usability, interoperability, and ease of implementation for the initial launch.

Question 42: Could you please confirm the compliance and accessibility standards (e.g., WCAG, ADA, Section 508) that your website must adhere to?

Answer: Yes, the platform should incorporate accessibility features that ensure equitable access across languages. At minimum, compliance with WCAG 2.1 accessibility standards will be required, including support for screen readers, captioning, and text alternatives. In addition, multilingual accessibility is an important consideration given the diverse communities expected to use the platform. This means that content, metadata, and user interfaces should be designed to accommodate non-English languages, and screen reader compatibility must extend beyond English text. While the initial scope may prioritize

English, the architecture should be built to enable future localization and translation, ensuring that multilingual accessibility can be expanded as the user base grows.

Question 43: The RFP references synthetic data tools (SDV, Synthea, CARLA). Should vendors provide a comparison and recommendation of tools, or implement a predefined stack?

Answer: There are no preferred synthetic data tools among SDV, Synthea, CARLA, or others. These are provided as illustrative examples, and respondents are free to propose alternative or additional tools they believe best meet the goals and requirements of the Data Commons.

Question 44: Technology Preferences: The RFP mentions tools like CKAN, Dataverse, SDV, Synthea, CARLA, AIF360, and Fairlearn. Are these stated as required solutions, preferred examples, or simply illustrative of the type of capability needed? Are respondents encouraged to propose alternative or additional technologies they believe are better suited?

Answer: The listed solutions are illustrative. Vendors should recommend appropriate solutions to meet project goals and needs.

Commons Credits

Question 45: What are the business rules for earning, spending, and tracking credits? Do you have a reference implementation or model?

Answer: Project scope includes designing and implementing the credits capability. Existing commons credit definition and context is available in DCC Requirements Section 7

Question 46: Any audit or security requirements for credit transactions?

Answer: Project scope includes designing and implementing the credits capability. Existing commons credit definition and context is available in DCC Requirements Section 7

Question 47: There is a deliverable “Explanation of planned industry engagement,” is this referencing the Commons Credits approach?

Answer: Planned industry engagement is a plan that will help launch the new DCC platform and includes developing targeted outreach, communications, and partnership strategies to raise awareness and encourage participation.

Question 48: Can MassTech please clarify what the expected data volumes are for: 1) Initial dataset hosting, 2) Synthetic data generation, 3) Commons credit transactions, 4) Use activity logs?

Answer: This requirement has not yet been defined. MassTech will consider recommendations from vendors and work collaboratively to establish an approach that aligns with best practices and project objectives.

Question 49: Can MassTech please clarify what the specific requirements are for the Commons Credits ledger regarding the following: Transaction processing speed, 2) Storage duration, 3) Integration points, and 4) Reporting requirements?

Answer: The Commons Credits ledger requirements included in the RFP are conceptual and intended to signal the need for a mechanism to track and manage usage across the platform. Specific requirements for transaction processing speed, storage duration, integration points, and reporting have not been prescribed. Vendors are encouraged to propose recommended approaches and capabilities in these areas, drawing from best practices and their experience implementing similar systems.

Question 50: Will vendors need to support credit trading between users?

Answer: No. Peer-to-peer credit trading between users is not in scope.

Question 51: What is the expected initial credit allocation model for Massachusetts agencies versus external researchers, and are there different pricing tiers planned for academic versus commercial usage?

Answer: At launch, the system will not include differentiated pricing tiers (e.g. academic vs. commercial). All credits will function uniformly as an incentive and resource-allocation mechanism within the DCC. Over time, as utilization patterns stabilize, the MA AI Hub may consider tiered allocation models, for example, subsidized or higher baseline credits for academic/nonprofit use, and paid or limited credits for commercial entities. Any such changes would be implemented after governance review and in compliance with state policy.

Question 52: Should the credits system integrate with existing Massachusetts procurement systems (COMMBUYS), and what are the accounting and invoicing requirements for cross-agency cost allocation?

Answer: The credits system is not required to integrate with COMMBUYS, nor to perform accounting or invoicing for cross-agency cost allocation.

Question 53: Are there specific sustainability metrics or carbon footprint tracking requirements that should be incorporated into the credit incentive structure mentioned in Section 7.4?

Answer: The DCC Requirements do not mandate specific sustainability metrics or carbon tracking at launch. However, because the MA AI Hub is committed to sustainable, green high-performance computing, the Commons Credits system should be designed with the flexibility to incorporate such metrics in the future. For example, credits could be adjusted to reward efficient compute usage (e.g., running jobs in off-peak, low-carbon periods) or to discourage wasteful resource consumption. At launch, the focus will be on dataset contributions, metadata enrichment, and fair usage of shared resources. Over time, sustainability metrics, such as carbon footprint tracking per workload or per dataset transfer, may be layered into the credit model, pending policy guidance and available measurement tools.

Question 54: Have you defined any success KPIs (i.e. dataset count, active users, credits utilized, etc.)?

Answer: KPI's include enabling impactful AI use cases (e.g., number of use cases, new products and services, research outputs), curating and utilizing high-value datasets, streamlining data sharing across sectors (e.g., exchanges, diversity of contributors, active usage), reducing barriers to AI development (e.g., user satisfaction), and ensuring equity and transparency (e.g., diversity of datasets and users). Vendors are encouraged to recommend additional KPIs that will help measure value and impact over time.

Question 55: Do you have an existing / preferred platform to support the commons credit capability?

Answer: No, there is no existing platform to support the commons credit capability, and MassTech does not have a preferred platform. Vendors are expected to propose recommended solutions based on best practices and alignment with the goals of the Data Commons.

Question 56: Will Commons Credits have integrations with services such as Hugging Face to pay for premium services on 3rd party platforms?

Answer: Not initially. At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 57: While the requirements document refers to the Commons Credits as a “Centralized Ledger”, would you consider a solution, such as bridge.xyz from Stripe, that allows you to deploy your own token on the blockchain? This may provide benefits such as currency conversion and acceptance at 3rd party providers such as Hugging Face while simplifying the token administration and management capabilities.

Answer: Yes, alternative approaches such as blockchain-based solutions may be considered. The reference to a “Centralized Ledger” in the requirements document is illustrative, and vendors are encouraged to propose innovative mechanisms—including tokenized or blockchain-enabled models—that can meet the goals of Commons Credits management, provided they align with compliance, security, and usability requirements.

Question 58: If you want to do a phased release, which data events should affect Commons Credits at launch (download, API call, curator time)?

Answer: MassTech is open to a phased release approach; however, establishing the Commons Credits capability is a priority. The specific data events that should affect credits at launch (e.g., downloads, API calls, curator time) have not yet been finalized and should be proposed by vendors as part of their recommended phasing strategy.

Question 59: Is a standard relational ledger sufficient, or do you want an append-only ledger with cryptographic audit features? You mention potential future blockchain—should we design for pluggable backends?

Answer: At launch, the system will use a cost-effective database and object storage backend but should be designed with a pluggable adapter layer, so that more advanced options (e.g., permissioned blockchain) can be adopted later without re-architecting. This balances audit integrity, usability, and long-term flexibility.

Question 60: How will credits be priced, funded (grants/purchase), and governed? Are credits considered a financial instrument (tax/accounting implications), and will they expire?

Answer: At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 61: The RFP requires the development of a "central ledger for tracking 'Commons Credits'". Could you provide additional details on the intended functionality of these credits? Is this a system for managing data access permissions, or is it a mechanism for tracking resource consumption and usage costs?

Answer: At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 62: Can you provide more detail on the intended use and structure of the “Commons Credits” ledger? Is this expected to be blockchain-based or a traditional database?

Answer: At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 63: Is the Commons Credits ledger envisioned as a blockchain-ready system or a traditional centralized database?

Answer: At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 64: Should the credits system integrate with real financial transactions (purchase of credits with currency) or only remain an internal incentive mechanism?

Answer: At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 65: Should credits also track compute resource consumption (GPU/CPU usage)?

Answer: Yes, compute usage should factor into commons credits.

Question 66: Should the ledger support smart contracts for automated enforcement of credit rules?

Answer: Yes. The Commons Credits ledger is expected to support smart contracts.

Question 67: Should credit transactions be auditable in real-time with APIs?

Answer: Yes.

Question 68: Should the ledger support interoperability with external systems (e.g., grant agencies providing credits)?

Answer: Not initially. At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 69: Should there be multi-currency support (credits ? USD exchange)?

Answer: Currency support is out of scope for this phase. At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 70: Should the Admin Dashboard include real-time usage analytics (dataset downloads, API calls, Commons Credits consumption) or is batch reporting sufficient?

Answer: Batch is sufficient for this phase.

Question 71: Should the dashboard include alerting features (e.g., low credit balance warnings, dataset access anomalies, expired consent tokens)?

Answer: Specific dashboard and reporting requirements have not yet been finalized. Vendors should propose a flexible reporting framework and include scope within the project to define detailed reporting needs in collaboration with MassTech. At a minimum, the platform should provide fundamental reports such as dataset usage and access metrics, system performance and uptime statistics, user activity and growth trends, and Commons Credits consumption summaries, with the ability to add more specialized dashboards as requirements evolve.

Question 72: "Commons Credits" System: Could you provide more detail on the envisioned purpose and mechanics of the "Commons Credits" ledger? Is this a mechanism for budgeting usage, incentivizing data contribution, or something else?

Answer: The Commons Credits ledger requirements included in the RFP are conceptual and intended to signal the need for a mechanism to track and manage usage across the platform. Specific requirements for transaction processing speed, storage duration, integration points, and reporting have not been prescribed. Vendors are encouraged to propose recommended approaches and capabilities in these areas, drawing from best practices and their experience implementing similar systems. See DCC Requirements Section 7 for additional details.

Question 73: To focus our approach, which 3-5 outcomes are the highest priority for the initial DCC launch (e.g., public portal live, # priority datasets onboarded, Commons Credits prototype live)? Are there acceptance criteria/KPIs for key deliverables?

Answer: Priority capabilities are listed in DCC Requirements Section 5 including, public data access via a website to the commons, curated data sets, synthetic data generation capability, etc.

Question 74: Please confirm initial earn/spend rules for Commons Credits at launch, whether real-currency purchases are in scope, and any legal/financial controls you require (e.g., issuance limits, reconciliation, refunds/chargebacks).

Answer: At launch, Commons Credits will operate as a closed-loop incentive system within the DCC, not a real-currency mechanism. Users will earn credits for contributions such as uploading datasets or enriching metadata and spend them on activities like higher-throughput API calls. Real-currency purchases are not in scope initially. Governance will include issuance limits, audit logs, and MA AI Hub oversight to prevent abuse, with the system designed to accommodate future extensions (e.g., monetary transactions) if pursued.

Question 75: Has Massachusetts Technology Collaborative explored the use of a public blockchain network to act as the underlying ledger for credit issuance and decentralized storage of large data sets?

Answer: No, this has not been explored but can be considered as part of a recommended solution

Community & Collaboration

Question 76: What marketing and outreach support is expected from MassTech to drive adoption across the Massachusetts innovation ecosystem?

Answer: A strategic engagement roadmap detailing how the bidder plans to involve, inform, and collaborate with industry customers of the DCC to drive relevance, contribution, and adoption is desired. The MA AI Hub will lead statewide marketing and outreach to drive adoption of the DCC in collaboration with the vendor. Vendors should be prepared to support by preparing technical collateral that the Hub can use in its outreach efforts.

Question 77: AI Use Cases - Because the platform is meant to 'accelerate AI development' and support 'research, government, startups, and industry,' what kinds of AI projects do you expect people to build—like chatbots, image recognition, or data analysis tools (for example, NLP for public records, computer vision for traffic data, or tabular modeling for healthcare outcomes)?

Answer: We anticipate a breadth of use case types to be developed on the platform including those referenced in the question.

Question 78: Synthetic Data - In light of the goal to 'enable synthetic data generation' and support tools like 'SDV, Synthea, CARLA,' do you expect users to create synthetic data using AI models—like generating fake patient records or traffic simulations (for example, using GANs, LLMs, or privacy-preserving techniques like differential privacy)?

Answer: Yes, the system should support synthetic data generation using AI models, see DCC Requirements Section 10

Question 79: Training, Rollout, and Adoption - Since the platform will 'train end users and administrators' and support a 'public launch,' will there be tools or environments for people to safely try out AI features—like testing prompts, generating synthetic data, or building small apps (for example, using sandbox environments, prompt playgrounds, or RAG-based assistants)?

Answer: Yes, the selected bidder should plan to set up a sandbox for this purpose.

Question 80: Will MassTech please clarify if there are specific data quality metrics required for synthetic data?

Answer: Specific data quality metrics for synthetic data have not yet been established. Vendors are encouraged to recommend appropriate metrics and evaluation methods, drawing on best practices to ensure that synthetic data is useful, reliable, and aligned with the goals of the Data Commons.

Question 81: In the DCC Requirements file in the table under section “3.0 Program Phases and Key Milestone Descriptions,” it says “At least one synthetic data tool configured and producing test outputs.” Is there a target for the number of synthetic data tools to be configured?

Answer: The reference to “at least one synthetic data tool configured and producing test outputs” in the requirements document is intended as a minimum illustrative target. Vendors may propose additional tools or approaches for synthetic data generation as part of their solution, based on their expertise and the value they believe can be delivered to the Data Commons.

Question 82: Are there specific synthetic data validation requirements or acceptable privacy-utility tradeoff thresholds that must be met for different data types (healthcare, transportation, etc.)?

Answer: Specific synthetic data validation requirements and privacy-utility tradeoff thresholds have not yet been established for different data types. Vendors are encouraged to propose recommended approaches, metrics, and safeguards based on best practices and compliance considerations (e.g., healthcare, transportation, or other sensitive domains).

Question 83: How should intellectual property rights be handled for synthetic data generated from Massachusetts agency datasets, and what attribution requirements exist?

Answer: Intellectual property requirements for different types of data have not yet been defined. These requirements will be determined as part of the project, and vendors are encouraged to recommend approaches and considerations based on best practices and compliance standards.

Question 84: For Synthetic Data Generation, does MassTech wish to white label this functionality or be a convenient pass through to tools that already exist?

Answer: MassTech is open to both approaches. Vendors may propose white-labeled synthetic data generation capabilities within the platform or convenient integrations with existing tools, with recommendations guided by usability, scalability, and alignment with the goals of the Data Commons.

Question 85: Which domain(s) do you want to prioritize for synthetic data pilots (e.g., health, mobility, climate)?

Answer: Top priority domains include life sciences, robotics, education, and healthcare

Question 86: Is the synthetic data generation capability expected to be developed from scratch or largely built by integrating existing tools (e.g., SDV, Synthea, CARLA)?

Answer: Synthetic data generation capabilities can be custom tools or based on existing tools. See DCC Requirements Section 10 for more details.

Question 87: What are the compliance benchmarks (e.g., HIPAA, GDPR, state privacy laws) we must demonstrate for synthetic data pipelines?

Answer: Specific compliance benchmarks for synthetic data pipelines have not yet been finalized. At a minimum, vendors should expect to align with applicable standards such as HIPAA, GDPR, and relevant Massachusetts state privacy laws, and are encouraged to recommend additional benchmarks and safeguards to ensure privacy, security, and responsible data use.

Question 88: Are there restrictions on subcontracting for specialized services (e.g., fairness audits, synthetic data tools)?

Answer: Subcontracting is permitted under this RFP, but any subcontractors must receive prior approval from the MA AI Hub before contract execution. Respondents are encouraged to identify proposed subcontractors in their submission if they are known, as this strengthens transparency and helps evaluators assess team qualifications. However, formal approval is only required before the contract is finalized, not at the time of proposal submission. The prime vendor will remain fully responsible for subcontractor performance and compliance.

Question 89: Are there preferred synthetic data tools among SDV, Synthea, and CARLA, or is the respondent free to propose alternatives?

Answer: There are no preferred synthetic data tools among SDV, Synthea, CARLA, or others. These are provided as illustrative examples, and respondents are free to propose alternative or additional tools they believe best meet the goals and requirements of the Data Commons.

Question 90: What level of training and capacity-building is expected (basic onboarding vs advanced fairness/synthetic data workshops)?

Answer: The specific level of training and capacity-building has not yet been defined. Vendors should propose a comprehensive approach that includes basic onboarding for users.

Question 91: Scope Prioritization: The scope of work is extensive. Should proposers assume all items under "Services Required" are of equal priority, or does MassTech have a prioritized list of capabilities (e.g., core catalog vs. synthetic data vs. fairness dashboard) for the initial launch?

Answer: MassTech has not established a formal prioritization of the items listed under "Services Required." Proposers should assume all are in scope but are encouraged to recommend phasing and prioritization (e.g., core catalog, synthetic data, fairness dashboard) based on their expertise and best practices for achieving a successful initial launch.

Question 92: Are there priority industries/sectors (e.g., health, energy, education) or use cases that MassTech intends to pilot first with the DCC?

Answer: Top priority domains include lifesciences, robotics, education, and healthcare

Data Content & Curation

Question 93: What is the expected scale (number and size) of datasets to be initially hosted and the projected growth?

Answer: The data sets listed in Appendix A represent potential data sets. The project scope includes finalizing a recommendation for which data sets should be sourced and included in the launch of the data commons.

Question 94: What are the anticipated data volumes, growth rates, and ingestion frequencies for each data source?

Answer: For the scope of this development and implementation, assume initial volumes are expected in the tens of TBs with ~30% annual growth, primarily ingested in batch or scheduled mode, while the architecture will be designed to scale to hundreds of TBs or more and support both batch and streaming ingestion as needs evolve.

Question 95: What is the estimated start date that vendors can use to inform the budget submitted?

Answer: Please use December 2, 2025 as the start date.

Question 96: Can MassTech please clarify if there's a desired deadline for the build of this Data Commons solution?

Answer: The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 97: In the DCC Requirements file in the table under section "1.0 Overview," please define "AICR."

Answer: In the DCC Requirements document, under section 1.0 Overview, the term "AICR" refers to the Artificial Intelligence Compute Resources initiative, a major investment by the Commonwealth to expand access to high-performance computing (HPC) through the Massachusetts Green High Performance Computing Center (MGHPCC). AICR is the state's shared HPC cluster designed to support AI research, innovation, and development. It will serve as the compute backbone for the Data Commons Collaborative.

Question 98: Which metadata standards are mandatory at launch (e.g., Dublin Core, DCAT, Schema.org), and what is the minimum completeness bar?

Answer: No specific metadata standards are mandated at launch. Vendors should recommend appropriate standards (e.g., DCAT, Schema.org, or others) and propose a practical minimum completeness bar that balances usability, interoperability, and ease of implementation for the initial launch.

Question 99: What is the expected timeline and process for adding new Massachusetts agencies or expanding to other New England states?

Answer: The RFP and DCC Requirements do not set a fixed timeline for onboarding additional Massachusetts agencies or expanding to other New England states. The platform architecture should be designed to support this expansion seamlessly as agreements are put in place.

Question 100: How should the DCC architecture accommodate emerging AI technologies (e.g., quantum computing integration mentioned in Section 7.7) and evolving regulatory requirements?

Answer: The DCC will use a modular, API-first, standards-aligned architecture so we can plug in new AI capabilities (e.g., quantum, agentic workflows) and adapt to evolving policy with minimal refactoring.

Question 101: What is your targeted Project Kickoff date?

Answer: 2025-12-02 00:00:00

Question 102: Do you have a target launch window?

Answer: The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 103: Are you targeting all features at launch or would you prefer a phased release to get the DCC live sooner? If you are considering a phased launch, have you identified your priority requirements for the initial launch? If you are considering a phased launch but have not defined scope, would you like us to include recommendations with a release roadmap?

Answer: A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 104: Have you identified a target budget?

Answer: Funding has been allocated for this project, but a specific budget ceiling is not being disclosed at this stage. We encourage vendors to submit competitive proposals with clear and transparent pricing, including breakdowns of one-time and ongoing costs as well as any optional features.

Question 105: What's the minimum metadata profile & standards (DCAT, Schema.org, Dublin Core) to support at launch?

Answer: No specific metadata standards are mandated at launch. Vendors should recommend appropriate standards (e.g., Dublin Core, DCAT, Schema.org, or others) and propose a practical minimum completeness bar that balances usability, interoperability, and ease of implementation for the initial launch.

Question 106: What's the depth of lineage and versioning needed at launch (dataset-level vs. column-level)? For example: Dataset-level lineage tracking → Captures the origin and transformations of the entire dataset. Example: showing that a climate dataset was ingested from MassGIS, processed in Databricks, and published in the AI Hub Commons on a certain date. It answers "Where did this dataset come from, and how has it changed as a whole?" Column-level lineage tracking → Provides a granular view of how each individual field/column was derived or transformed. Example: tracing the "Average Temperature" column back to raw sensor data, including formulas or transformations applied. It answers "How was this specific attribute produced, and from what source fields?"

Answer: The required depth of lineage and versioning at launch has not yet been finalized. At a minimum, dataset-level lineage tracking will be needed to ensure transparency around dataset origins and transformations, while column-level lineage may be incorporated in later phases as part of an enhanced capability. Vendors are encouraged to propose approaches that balance usability, scalability, and cost, while providing a roadmap for evolving from dataset-level lineage at launch to more granular lineage over time.

Question 107: If the dataset origination source doesn't support CSV, JSON, or Parquet, will the DCC be responsible for the data conversion?

Answer: If a dataset's originating source does not support common formats such as CSV, JSON, or Parquet, the approach for data conversion will need to be addressed as part of the project. Vendors are encouraged to recommend strategies and tools for handling conversion and standardization to ensure datasets are accessible, interoperable, and usable within the Data Commons.

Question 108: Will there be oral presentations or a best-and-final offer round, and when would they occur relative to the 9/23 due date?

Answer: MA-AI Hub leaves open the possibility of oral presentations.

Question 109: What is the target timeline from award through public launch? The RFP lists deliverables and the DCC doc lists phases but no dates. Please confirm expected start date and phase durations.

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 110: Should the vendor propose a single tool (e.g., SDV, Synthea) or build a toolkit integration layer supporting multiple?

Answer: There are no preferred synthetic data tools among SDV, Synthea, CARLA, or others. These are provided as illustrative examples, and respondents are free to propose alternative or additional tools they believe best meet the goals and requirements of the Data Commons.

Question 111: Data Sensitivity & Tiering: The RFP mentions "dataset tiering." Can you elaborate on the anticipated data classification tiers (e.g., public, internal, restricted) and the specific security and access control protocols envisioned for each tier?

Answer: Dataset tiering has not yet been defined; vendors should recommend an approach.

Question 112: Compliance Frameworks: Are there specific compliance frameworks beyond general best practices that the DCC must adhere to? (e.g., NIST, CIS Controls, CJIS, HIPAA, or specific Massachusetts data security laws like CMR 17.00).

Answer: Specific compliance frameworks are not yet defined. Vendors are encouraged to make recommendations based on expertise industry standards.

Question 113: Is there a target launch date for the DCC or an anticipated timeline for the build?

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 114: Has a target Go-Live date been established for the DCC?

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 115: Is there flexibility in the budget structure to support a phased or modular proposal (e.g., base build vs. optional enhancements)?

Answer: Yes, a phased approach is desired.

Question 116: Timeline: What is the anticipated duration for the project?

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Data Governance & Policy

Question 117: Consent needs are described as tokenized, revocable, and time-bound. Please confirm whether this is in scope for phase one, and whether consent applies to public/open datasets or only restricted data?

Answer: Yes, tokenized, revocable, and time-bound consent management is in scope for phase one of the DCC. Consent applies primarily to restricted datasets where contributors need control over access and revocation.

Question 118: For consent management, should vendors implement a tokenized dynamic consent system or is a simpler consent capture acceptable?

Answer: Yes, tokenized, revocable, and time-bound consent management is in scope. Consent applies primarily to restricted datasets where contributors need control over access and revocation.

Question 119: The budget template implies a 4-year period of performance. Is there a 4-year period of performance?

Answer: The budget template does not specify or imply a 4-year period of performance. There is a post-contract section for invoicing that has 4 columns, but the periods within each column are not defined.

Question 120: Will the awardee of this RFP be expected to work with all datasets in Appendix A of the "DCC Requirements.pdf" during the period of performance?

Answer: The data sets listed in Appendix A represent potential data sets. The project scope includes finalizing a recommendation for which data sets should be sourced and included in the launch of the data commons.

Question 121: Are there multilingual accessibility requirements (e.g., screen reader support for non-English content)?

Answer: Yes, the platform should incorporate accessibility features that ensure equitable access across languages. At minimum, compliance with WCAG 2.1 accessibility standards will be required, including support for screen readers, captioning, and text alternatives. In addition, multilingual accessibility is an important consideration given the diverse communities expected to use the platform. This means that content, metadata, and user interfaces should be designed to accommodate non-English languages, and screen reader compatibility must extend beyond English text. While the initial scope may prioritize English, the architecture should be built to enable future localization and translation, ensuring that multilingual accessibility can be expanded as the user base grows.

Question 122: Given the potential impact of forthcoming Q&A responses on vendor proposals, would MassTech consider extending the submission deadline by a minimum of two weeks? This extension would ensure vendors can thoroughly incorporate any clarifications or changes stemming from the Q&A process into their final proposals.

Answer: Yes, the proposal deadline has been extended by 2 weeks.

Question 123: The RFP appears to focus primarily on consulting services, as reflected in both the title and the budget template. However, Section 2 (Services Required) includes significant technical implementation requirements such as building infrastructure, implementing metadata management tools, and developing synthetic data capabilities. Could MassTech clarify if they would like vendors to respond with a technical solution?

Answer: Yes, the RFP scope includes technical design, implementation and launch.

Question 124: In Appendix C Services Budget Template (Budget Invoice Reporting tab), what does "MTC Funding Requested" mean?

Answer: "MTC Funding Requested" refers to the portion of the project cost that the Respondent is requesting MassTech to fund under this contract. In this case, the amount will likely be the same as the Total Project Cost. There are situations (more common in grants than in service contracts) where the Respondent contributes its own resources or funding, and in those cases the "MTC Funding Requested" would be less than the Total Project Cost.

Question 125: In Appendix C Services Budget Template (Budget Invoice Reporting tab), what elements do you consider "Direct Labor Fringe Cost" and "General & Administrative Expense/Overhead"?

Answer: “Direct Labor Fringe Cost” generally refers to the employer’s cost of employee benefits associated with direct project labor. This may include items such as payroll taxes, health insurance, retirement contributions, and paid leave.

"General & Administrative Expense/Overhead" refers to indirect costs that support the overall operations of the vendor but are not directly attributable to a specific project activity. Examples may include office space, utilities, administrative staff, and general business expenses.

Vendors may either itemize these costs separately in the budget template or propose fully-loaded labor rates that already include fringe and overhead.

Question 126: For post-launch requirements (i.e. system maintenance, performance updates), is MassTech expecting the awarded vendor to perform these services, or will you be engaging another vendor to support or utilize in-house resources?

Answer: MassTech anticipates that the awarded vendor will provide initial system stabilization, documentation, and knowledge transfer to ensure a smooth transition after launch. Long-term operations, such as system maintenance, performance tuning, and future feature expansion may be supported either by the original vendor, by a separate contracted provider, or by in-house resources, depending on program needs and budget.

Question 127: Can MassTech please define the following items related to testing and quality assurance: 1) Acceptance criteria for each delivery, 2) User acceptance testing scope and duration, 3) Performance testing benchmarks, and 4) Security testing requirements?

Answer: Detailed requirements for acceptance criteria, user acceptance testing (UAT), performance benchmarks, and security testing will be defined collaboratively during project initiation and planning. Vendors are expected to propose recommended approaches, frameworks, and best practices for these testing and quality assurance activities as part of their response, with the final scope and criteria to be agreed upon with MassTech during execution.

Question 128: Will there be subject matter experts from MassTech available to support the awardee during implementation?

Answer: Yes.

Question 129: Are there specific requirements for handling proprietary algorithms or models developed using DCC resources, including commercialization rights and revenue sharing?

Answer: Intellectual property generated by users of the DCC (e.g., models, algorithms, or applications) will remain with those users or their institutions, subject to any agreements they may have with their funders or collaborators. The DCC itself will not assert ownership rights over derivative works, but it will enforce Terms of Use and ethical guidelines to ensure compliance with data-sharing agreements and licensing conditions.

Question 130: What monitoring and alerting capabilities are required for performance tracking and proactive issue resolution?

Answer: Specific monitoring and alerting requirements have not yet been defined. Vendors should propose best-practice approaches for performance tracking, system health monitoring, and proactive issue resolution.

Question 131: Regarding the guidance to provide alternatives that provide "more cost-effective performance", will you consider responses that utilize our offshore employees in addition to our US based team? We can deliver with fully onshore resources or using a hybrid model depending on your needs.

Answer: The MA AI Hub requires that key functions involving governance, security, compliance, and stakeholder engagement be performed by U.S. based personnel. Vendors may propose hybrid delivery models that include offshore team members for defined roles such as software development or testing, provided this does not compromise data residency, security, or responsiveness. Proposals should clearly explain the division of onshore vs. offshore roles and how communications and compliance will be maintained. Fully onshore delivery models will also be considered and may be viewed as lower risk. As it pertains to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 132: Will you have regular content updates such as announcements, blogs, that will require a Content Management Capability?

Answer: Yes, regular content updates such as announcements and blogs are expected. However, the approach for managing this content has not been predefined, and vendors are encouraged to recommend appropriate content management capabilities as part of their solution.

Question 133: Will MassTech be the primary for procuring any license or hosting platform or will that be a pass-through cost from the selected vendor?

Answer: MassTech will serve as the primary for procuring any required licenses or hosting platforms. These costs should not be treated as vendor pass-through expenses in proposals.

Question 134: We have noted your policy on travel and reimbursable expenses. What are your desired onsite expectations? Do you want team members onsite daily, on a defined regular cadence, or in support of major meetings and milestones?

Answer: Onsite expectations are in support of milestones and meetings as needed.

Question 135: Are you requiring all resources to be US-based, or can offshore resources be used as well?

Answer: The MA AI Hub requires that key functions involving governance, security, compliance, and stakeholder engagement be performed by U.S.-based personnel. Vendors may propose hybrid delivery models that include offshore team members for defined roles such as software development or testing, provided this does not compromise data residency, security, or responsiveness. Proposals should clearly explain the division of onshore vs. offshore roles and how communications and compliance will be maintained. Fully onshore delivery models will also be considered and may be viewed as lower risk.

Question 136: Should the architecture support containerized deployment (Docker/Kubernetes) for modular services?

Answer: Yes

Question 137: Should metadata ingestion support auto-harvesting from external sources (e.g., harvesting CKAN/Dataverse catalogs) or only manual curation?

Answer: Requirements are not prescriptive; the solution should support both auto-harvesting and manual curation. Vendors are encouraged to propose connectors for harvesting external catalogs (e.g., CKAN/Dataverse), with scheduling, schema mapping, and de-duplication, alongside workflows for manual review/enrichment to ensure quality and governance. An MVP may prioritize manual curation with targeted harvesters, with a roadmap to expand automated ingestion based on demand.

Question 138: Will MassTech supply the initial curated dataset list, or is the vendor responsible for sourcing & vetting datasets from external/public repositories?

Answer: The DCC Requirements document includes examples of potential datasets, but a finalized curated list has not yet been defined. Vendors should expect to work collaboratively with MassTech to refine the initial dataset strategy, which may include sourcing and vetting datasets from external or public repositories.

Question 139: Should the vendor onboard all datasets listed in Appendix A, or only a subset for Phase 1?

Answer: The data sets listed in Appendix A represent potential data sets. The project scope includes finalizing a recommendation for which data sets should be sourced and included in the launch of the data commons.

Question 140: Should vendor conduct penetration testing & security certification or will state security teams handle this?

Answer: The requirement for penetration testing and security certification has not yet been finalized. Vendors should propose recommended approaches for conducting these activities, with the understanding that final responsibilities will be determined in coordination with MassTech and applicable stakeholders during the project.

Question 141: Should encryption use state-approved KMS (Key Management Service) or can vendor-provided solutions suffice?

Answer: There is no current mandate for Key Management Service (KMS). Vendors may propose their own secure, standards-compliant KMS solutions or recommend use of state-approved services, with the final approach to be determined in collaboration with MassTech during the project.

Question 142: Should MFA support FIDO2/WebAuthn hardware keys or just TOTP apps (Google Authenticator, Authy)?

Answer: Detailed MFA requirements have not yet been defined. Vendors should propose secure authentication solutions—whether TOTP apps, FIDO2/WebAuthn hardware keys, or a combination—that best meet security and usability needs.

Question 143: Should multi-language support (Spanish, Chinese, etc.) be part of Phase 1 accessibility?

Answer: Yes, the platform should incorporate accessibility features that ensure equitable access across languages. At minimum, compliance with WCAG 2.1 accessibility standards will be required, including support for screen readers, captioning, and text alternatives. In addition, multilingual accessibility is an important consideration given the diverse communities expected to use the platform. This means that content, metadata, and user interfaces should be designed to accommodate non-English languages, and screen reader compatibility must extend beyond English text. While the initial scope may prioritize English, the architecture should be built to enable future localization and translation, ensuring that multilingual accessibility can be expanded as the user base grows.

Question 144: Is the vendor expected to support post-launch operations (SLA, bug fixes, enhancements), or is handoff at launch sufficient?

Answer: MassTech anticipates that the awarded vendor will provide initial system stabilization, documentation, and knowledge transfer to ensure a smooth transition after launch.

Question 145: Should the platform's UI support multi-language capabilities at launch (English + others), or is multilingual support expected only in future phases?

Answer: Yes, the platform should incorporate accessibility features that ensure equitable access across languages. At minimum, compliance with WCAG 2.1 accessibility standards will be required, including support for screen readers, captioning, and text alternatives. In addition, multilingual accessibility is an important consideration given the diverse communities expected to use the platform. This means that content, metadata, and user interfaces should be designed to accommodate non-English languages, and screen reader compatibility must extend beyond English text. While the initial scope may prioritize English, the architecture should be built to enable future localization and translation, ensuring that multilingual accessibility can be expanded as the user base grows.

Question 146: Are there specific load time/speed performance, and application uptime benchmarks the site should achieve?

Answer: This has not been defined. Vendors should recommend policies based on best practices.

Question 147: What is the expected duration and scope of post-launch support and maintenance?

Answer: The duration and scope of post-launch support have not yet been defined and should be included as part of the vendor's recommended project scope.

Question 148: What are the expected support hours and days for system maintenance and assistance (e.g., 24/7 support, business hours, or specific timeframes).

Answer: Ongoing 24/7 system support is not in scope for this RFP. Vendors are expected to provide support during the design, build, testing, and initial rollout phases, including documentation, training, and assistance with launch. Future operational support models is not in scope.

Question 149: How will the success of the Data Commons Collaborative (DCC) be measured in its first year post-launch?

Answer: KPI's include enabling impactful AI use cases (e.g., number of use cases, new products and services, research outputs), curating and utilizing high-value datasets, streamlining data sharing across sectors (e.g., exchanges, diversity of contributors, active usage), reducing barriers to AI development (e.g., user satisfaction), and ensuring equity and transparency (e.g., diversity of datasets and users). Vendors are encouraged to recommend additional KPIs that will help measure value and impact over time.

Question 150: Is there an existing technical / technological / vendor footprint we need to account for, or start anew?

Answer: Start new, no existing technical footprints exist

Question 151: Are you open to a hybrid delivery model with a mix of offshore and onshore resources?

Answer: The MA AI Hub requires that key functions involving governance, security, compliance, and stakeholder engagement be performed by U.S.-based personnel. Vendors may propose hybrid delivery models that include offshore team members for defined roles such as software development or testing, provided this does not compromise data residency, security, or responsiveness. Proposals should clearly explain the division of onshore vs. offshore roles and how communications and compliance will be maintained.

Fully onshore delivery models will also be considered and may be viewed as lower risk. As it pertains specifically to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 152: Performance Metrics: What performance metrics or success criteria will be used to evaluate the effectiveness of the consulting services once implemented?

Answer: Success for the DCC will be measured by both ecosystem engagement and user experience. Key outcomes include securing diverse partners across academia, government, startups, and industry; formalizing collaborations through MOUs; and convening events that drive adoption and data contributions. Performance metrics will focus on user satisfaction and usability survey results as well as measurable collaboration indicators such as the number of data exchanges between DCC users and the diversity of domains represented in those exchanges.

Integration & Interoperability

Question 153: Metadata & Dataset Curation - As the platform is expected to 'ensure metadata completeness and FAIR data principles,' will users be able to search and filter datasets based on how useful they are for AI—like finding data that's ready for training or labeled for specific tasks (for example, using schema.org, DCAT, or embedding-based semantic search)?

Answer: The system should support data discovery capabilities (see DCC Requirements Section 8). Including AI ready data sets is in scope.

Question 154: Should the solution support data access rate limiting, quotas, or API key management?

Answer: Yes. For a multi-tenant, API-based platform, we should build in rate-limiting, quotas, and API key/token management.

Question 155: Can Massachusetts Technology Collaborative confirm that a rate-based contract will be awarded in response to this RFP?

Answer: Massachusetts Technology Collaborative anticipates structuring the award in a way that aligns with best practices for both value and accountability. While a rate-based contract is an available vehicle, we are also open to fixed-price or hybrid approaches where appropriate. Our priority is to minimize the risk of budget overruns while ensuring timely high-quality delivery. We encourage respondents to propose the pricing model (rate-based, fixed price, or blended) that best matches their approach, with clear assumptions, deliverables, and mechanisms for cost control.

Question 156: Can Massachusetts Technology Collaborative confirm that a rate-based contract will be awarded in response to this RFP, and that bidders may submit a budget built with loaded rates (inclusive of indirect costs and fee)?

Answer: Massachusetts Technology Collaborative anticipates structuring the award in a way that aligns with best practices for both value and accountability. While a rate-based contract is an available vehicle, we are also open to fixed-price or hybrid approaches where appropriate. Our priority is to minimize the risk of budget overruns while ensuring timely high-quality delivery. We encourage respondents to propose the pricing model (rate-based, fixed price, or blended) that best matches their approach, with clear assumptions, deliverables, and mechanisms for cost control.

Question 157: Can Massachusetts Technology Collaborative please confirm that bidders may leave certain fields blank in the pricing template if those fields do not apply (e.g., fringe)?

Answer: Fields that are not needed for a bidder submission can be left blank

Question 158: Are there any eligibility restrictions regarding who can respond to this RFP?

Answer: No.

Question 159: Will the contractor be responsible for storing any data in the contractor's computing environment, or will all data be handled within a Massachusetts Technology Collaborative computing environment? If the contractor will be storing data, will all data be publicly available data, or will any data contain PII or PHI?

Answer: It is anticipated that through this project the required data systems will be designed, developed and implemented to house data and connect to data sets as part of the data commons capability. Contractor data storage is out of scope.

Question 160: The "DCC Requirements.pdf" has a table titled "3.0 Program Phases and Key Milestone Descriptions". Are these the milestones expected to be completed by the awardee of this RFP?

Answer: The phases and milestones outlined in the DCC Requirements.pdf (section 2.0) are provided as a notional framework to illustrate the anticipated scope and progression of the program. Vendors are encouraged to propose alternative or refined phasing and milestones as part of their response, based on their expertise, methodology, and recommended approach to achieving the program's goals.

Question 161: The RFP Services Budget Template (Budget Guidelines tab) states that "These guidelines apply to MassTech grants that require Participants to incur expenditures in line with approved budgets within their contracts." Is this the correct pricing workbook for vendors to complete, given that the solicitation is for professional services to implement a technical solution - and not a grant-based project?

Answer: Yes, this is the correct pricing workbook. While the Budget Guidelines tab references grants, these guidelines are applicable to both grants and services contracts.

Question 162: Is AI-driven discovery allowed? Will it include the Commonwealth's recent "Usage of Generative AI" contract language?

Answer: Yes. AI-driven discovery features (e.g., intelligent dataset recommendations, semantic search, or metadata enrichment) are allowable within the DCC as long as they are implemented in compliance with Commonwealth policy. The Commonwealth of Massachusetts has recently adopted standard contract language on the usage of Generative AI, which governs how such tools may be deployed to ensure privacy, security, accuracy, and ethical safeguards. Any use of AI-driven discovery within the platform will be aligned to those provisions, meaning outputs must be transparent, non-determinative, and auditable.

Question 163: What are the specific performance benchmarks beyond <2s response times, particularly for complex queries across federated datasets or AI model training workloads?

Answer: Specific performance benchmarks have not yet been established. Requirements for complex queries across federated datasets and AI model training workloads will be determined as part of the project, and vendors are encouraged to propose recommended benchmarks and scalability approaches based on best practices.

Question 164: Do you have a preference or requirement to utilize our Massachusetts based team members above team members in other states?

Answer: The MA AI Hub requires that key functions involving governance, security, compliance, and stakeholder engagement be performed by U.S. based personnel. Vendors may propose hybrid delivery models that include offshore team members for defined roles such as software development or testing, provided this does not compromise data residency, security, or responsiveness. Proposals should clearly explain the division of onshore vs. offshore roles and how communications and compliance will be maintained. Fully onshore delivery models will also be considered and may be viewed as lower risk. As it pertains to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 165: Do you plan on using the selected Data Management portal (CKAN, Dataverse, etc.) for the complete web experience or do you have a chosen/preferred web platform (WordPress, Drupal, Adobe, Optimizely, Sitecore, etc.)? If you prefer a CMS consideration, should we include platform selection in the response scope?

Answer: No final decisions have been made regarding data management portal solutions. Vendors should recommend an approach that aligns with best practices and project objectives.

Question 166: Will the look and feel of the AI Hub match the broader MassTech branding? If yes, will you be providing Brand Guidelines? If no, is that a requirement that should be included in the scope of our response?

Answer: The MA AI Hub brand guidelines will be available during project execution.

Question 167: Will MassTech accept a reasonable limitation of liability clause in the contract?

Answer: Proposed contract provisions will be reviewed as part of the contracting phase.

Question 168: Will bidders be required to accept the contract "as is" with no negotiation, or is there room for contractual modifications?

Answer: Proposed contract provisions will be reviewed as part of the contracting phase.

Question 169: What are the insurance requirements and indemnification obligations under the contract?

Answer: Please review MassTech's procurement webpage for examples of contract language used in various agreement templates. Specific language will be discussed as part of the contracting phase.

Question 170: The RFP lists Dataverse as an example of a metadata management tool. Could you clarify if the Massachusetts AI Hub has any existing data infrastructure or tools currently in use that the solution would need to integrate with? Is there a preference for an on-premises, cloud-native, or hybrid deployment model for the Data Commons Collaborative?

Answer: The Massachusetts AI Hub does not currently have existing data infrastructure or metadata management tools in place that the solution must integrate with. There is no preference at this stage regarding deployment model; solutions may be proposed as on-premises, cloud-based, or hybrid, with vendors encouraged to recommend the approach that best balances scalability, security, cost, and long-term sustainability.

Question 171: Is there a preference between CKAN and Dataverse as a metadata tool, or should the respondent propose based on best fit?

Answer: There is no preference between CKAN and Dataverse as a metadata tool. These are provided as illustrative examples, and respondents should propose the solution they believe is the best fit based on scalability, interoperability, usability, and alignment with the goals of the Data Commons.

Question 172: Is pre-approval of any subcontractors required at the time of submission, or only prior to contract execution?

Answer: Subcontracting is permitted under this RFP, but any subcontractors must receive prior approval from the MA AI Hub before contract execution. Respondents are encouraged to identify proposed subcontractors in their submission if they are known, as this strengthens transparency and helps evaluators assess team qualifications. However, formal approval is only required before the contract is finalized, not at the time of proposal submission. The prime vendor will remain fully responsible for subcontractor performance and compliance.

Question 173: Should semantic search (NLP-based dataset search) be supported in Phase 1?

Answer: Yes, semantic search should be available

Question 174: How much emphasis should be given to Massachusetts-specific datasets vs global datasets?

Answer: Both Massachusetts-specific and global datasets are important to the Data Commons. Vendors should recommend a balanced approach that maximizes impact.

Question 175: How long should audit logs be retained (docs mention 3+ years — confirm exact requirement)?

Answer: The current requirement is to retain audit logs for at least three years. This duration may be adjusted based on a more detailed scoping and compliance review conducted after the project begins.

Question 176: Should there be an API gateway for external integration (rate limits, API keys, monitoring)?

Answer: Yes

Question 177: The RFP suggests that pricing to be provided using Attachment C – Budget Template, but the attachment template itself appears to reference use post-contract. Could you please confirm how respondents should submit the pricing?

Answer: The Total Project Budget should be entered in Column J of the Attachment C - Budget Template. Columns K-T are provided for post-contract invoicing and are not to be completed as part of the proposal submission.

Question 178: DCC Requirements Document: The RFP references an attached "DCC Requirements.docx." Was this intended to be part of the released RFP package? If so, we do not see it attached and request it be provided to ensure our proposal is fully responsive.

Answer: The DCC Requirements document can be found at the following address: https://masstech.org/sites/default/files/2025-08/DCC%20Requirements.pdf

Question 179: Budget Template (Attachment C): The RFP references a Budget Template as an Excel spreadsheet. Was this intended to be part of the released RFP package? We do not see it attached and request it be provided to ensure our cost submission is formatted correctly.

Answer: The Budget Template was posted along with the RFP on the MassTech website: https://masstech.org/request-proposals-ai-hub-data-commons-collaborative-consulting-services

Question 180: Travel Assumptions: The RFP states travel will be reimbursed at the IRS mileage rate. Should we assume all engagement activities are expected to be conducted remotely, or is some amount of on-site work in Massachusetts anticipated? This will impact cost assumptions.

Answer: In person meetings may be required as needed or for major milestones.

Question 181: Do you require specific differential-privacy parameters (epsilon/delta) or is qualitative privacy validation sufficient initially? Please also confirm acceptable validation methods (e.g., membership inference tests, DP accounting reports).

Answer: There is currently no mandate for specific differential-privacy parameters (e.g., epsilon/delta) or prescribed validation methods. Vendors should recommend appropriate solutions and validation approaches—such as membership inference tests or DP accounting reports—as part of their proposal and project scope.

Question 182: What assumptions should vendors make regarding ongoing support and maintenance after the initial launch? Will that be re-procured, transitioned to the Massachusetts Technology Collaborative, or expected as part of this scope in the form of Managed Services?

Answer: MassTech anticipates that the awarded vendor will provide initial system stabilization, documentation, and knowledge transfer to ensure a smooth transition after launch.

Question 183: If this is a new Contract, what is the annual Budget for this?

Answer: Massachusetts Technology Collaborative anticipates structuring the award in a way that aligns with best practices for both value and accountability. While a rate-based contract is an available vehicle, we are also open to fixed-price or hybrid approaches where appropriate. Our priority is to minimize the risk of budget overruns while ensuring high-quality delivery. We encourage respondents to propose the pricing model (rate-based, fixed price, or blended) that best matches their approach, with clear assumptions, deliverables, and mechanisms for cost control.

Question 184: Is this contract intended to be awarded to a single vendor or to multiple vendors?

Answer: Proposers should assume a prime-vendor model (with approved subs as needed) while understanding that MassTech may, at its discretion, make multiple awards or structure work across vendors.

Question 185: Evaluation Criteria: Are there specific metrics that you can share for each of the evaluation criteria in section 4.2 of the RFP (i.e. relative weighting/total points)?

Answer: The criteria identified are provided as a comprehensive set of evaluation factors and are not intended to represent a prioritized or weighted list. The Commonwealth will apply these criteria in a balanced and holistic manner to assess the qualifications, capabilities, and overall value of each submitter in determining the best interest of the Data Commons initiative.

Metadata & Lineage

Question 186: Agentic Workflows - Since the platform will offer 'API-based data access' and serve many types of users, do you expect it to also support building smart AI tools that can automatically connect to different data and services—like using LangChain or DSPy for agentic workflows (for example, building AI agents using MCP, or chaining tools for IoT or legal research assistants)?

Answer: Yes. The Data Commons will support agentic workflows by enabling API-based access that can be connected to frameworks like LangChain or DSPy, allowing users to build smart AI tools and agents.

Question 187: Network and Access - As the platform will offer 'API-based data access' and serve multiple user groups, will there be requirements for high-throughput, secure networking—like multi-tenant access, VPN integration, or segmented access control (for example, using software-defined networking or secure virtual private networks)?

Answer: The solution should be designed with the following principles:

Multi-tenant Access & Role Segmentation - Role-based access control (RBAC) and dataset tiering, as called for in the requirements, will ensure that public, partner, and restricted datasets are clearly segmented. Tenant isolation will be enforced at the network and API gateway layers, with per-tenant quotas, authentication, and logging.

Software-Defined Segmentation-Network segmentation will be implemented logically (e.g., by project, dataset tier, or environment) to reduce risk and align with the governance framework. SDN or namespace-based policies will separate workloads (e.g., discovery portal, metadata services, hosted datasets, synthetic data generators).

Scalable Throughput-Initial performance targets will support API and bulk dataset transfers with room to scale. Streaming and high-volume use cases will be prioritized for efficient egress, while routine catalog and metadata queries will remain low-latency.

Audit & Compliance - Access requests, dataset downloads, and API calls will be logged and auditable to meet the platform's security and ethical use requirements.

Question 188: How do you envision the evolution of the platform over the next 1 to 3 years? Any long term data, AI or digital transformation goals we should consider?

Answer: Over the next 1–3 years, we envision the Data Commons evolving from a baseline catalog and secure API access point into a robust platform that actively fuels Massachusetts'

AI ecosystem. In the first year, the focus will be on establishing core infrastructure: dataset cataloging, FAIR metadata, role-based access, audit trails, and synthetic data capabilities. The platform should expand to include richer analytics (e.g., dataset quality scores, usage dashboards), and integrated bias/fairness monitoring. By year three, the platform could enable streaming data ingestion. In the longer term, the Data Commons should be viewed as a strategic digital transformation asset for Massachusetts, supporting not only equitable access to AI-ready data, but also enabling workforce development, startup acceleration, and public-sector innovation. This trajectory positions the platform not as a static repository, but as a living, adaptive commons that grows with both user needs and the pace of AI innovation.

Question 189: Is the goal to centralize all data to a single location or will federated access to data be preferred?

Answer: Federated access to data is required

Question 190: Have all data sources been identified and inventoried?

Answer: No, potential data sources have been identified (see DCC Requirement Appendix A). Finalizing the initial data sets and data sourcing are in scope for the project.

Question 191: Have business owners for each data source been identified?

Answer: Business owners for each data source have not yet been identified and is in scope for the project.

Question 192: Are there data locality, residency, or sovereignty requirements that should be considered in the architecture?

Answer: By default, all data (including logs, backups, and model artifacts) will be stored and processed in U.S. regions with an East-coast preference and the option to pin workloads to MGHPCC (on-prem/private) for sensitive use cases or as appropriate in cloud service. As it pertains specifically to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 193: Are you or do you have a data governance or MDM platform currently?

Answer: MassTech does not currently operate a centralized data governance or master data management (MDM) platform for this initiative. The expectation is that the selected solution will provide governance capabilities—including role-based access controls, audit trails, and metadata management—appropriate to support the Data Commons. We encourage proposals to describe how their platforms can integrate with existing state IT standards and, if applicable, interoperate with commonly used MDM or governance tools to ensure scalability and interoperability.

Question 194: Will data sets be curated or designed for specific use cases before being available to the userbase or is it direct access to the data stores listed?

Answer: The Data Commons is expected to support a hybrid approach. Some datasets will be curated and structured to enable specific research or use cases, while others may be made available in raw or semi-structured form with metadata and provenance information to support broader exploration. Proposals should describe how their platforms accommodate both curated and raw data access, while ensuring appropriate permissions, transparency, and ease of discovery for end users.

Question 195: Can MassTech confirm that awardees will not be responsible for data licensing and acquisition costs?

Answer: Yes. MassTech confirms that awardees will not be responsible for data licensing or acquisition costs; these costs will be addressed separately.

Question 196: Can MassTech please specify what the data retention requirements are for different types of data?

Answer: Data retention requirements for different types of data have not yet been defined. These requirements will be determined as part of the project, and vendors are encouraged to recommend approaches and considerations based on best practices and compliance standards.

Question 197: In the DCC Requirements file in section "5.0 Key Functional Requirements - External Data Sets and Data Capabilities," how many initial datasets need to be integrated? Are specific initial datasets required to be integrated?

Answer: There is no set number of initial datasets required to be integrated, and no specific datasets are mandated. The examples provided in the requirements are illustrative; vendors are encouraged to recommend an initial set of datasets that would best demonstrate the platform's capabilities and serve priority use cases.

Question 198: What specific Massachusetts state data residency and sovereignty requirements must be met, and are there restrictions on cross-border data processing or storage that would impact multi-cloud deployment strategies?

Answer: By default, all data (including logs, backups, and model artifacts) will be stored and processed in U.S. regions with an East-coast preference and the option to pin workloads to MGHPCC (on-prem/private) for sensitive use cases or as appropriate in cloud service. As it pertains specifically to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 199: Beyond HIPAA compliance mentioned in Section 14.1, what other Massachusetts-specific data protection regulations (e.g., Massachusetts Data Protection Regulation 201 CMR 17.00) must be incorporated into the governance framework?

Answer: There are not currently identified compliance frameworks, vendors are invited to recommend requirements based on project objectives and goals.

Question 200: Are there specific geographic requirements for data center locations within Massachusetts or New England for latency optimization and compliance purposes?

Answer: By default, all data (including logs, backups, and model artifacts) will be stored and processed in U.S. regions with an East-coast preference and the option to pin workloads to MGHPCC (on-prem/private) for sensitive use cases or as appropriate in cloud service. As it pertains to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 201: What specific community stakeholder groups (beyond academic institutions) should be included in the governance framework, and are there requirements for public comment periods on policy changes?

Answer: A governance committee is in place and will participate actively in the project. Requirements for public comment periods on policy changes have not yet been defined and will be determined as part of the project, with vendors encouraged to recommend approaches that support inclusivity and accountability.

Question 202: What are the data retention and deletion requirements for different data types, particularly for research projects with limited-term access agreements?

Answer: Data retention requirements for different types of data have not yet been defined. These requirements will be determined as part of the project, and vendors are encouraged to recommend approaches and considerations based on best practices and compliance standards.

Question 203: Are there specific requirements for security clearances or background checks for personnel accessing sensitive datasets?

Answer: Specific requirements for security clearances or background checks for personnel accessing sensitive datasets have not yet been defined. These requirements will be determined as part of the project, and vendors are encouraged to propose recommended approaches consistent with best practices for handling sensitive data.

Question 204: Beyond DICOM, NIfTI, and mp4 mentioned in the use cases, what other data formats must be supported for Massachusetts-specific datasets (e.g., GIS formats for MassGIS integration)?

Answer: Beyond DICOM, NIfTI, and mp4, no additional formats are formally mandated at this stage. However, the Data Commons is expected to support a broad range of formats relevant to Massachusetts-specific datasets and cross-sector AI research. Examples include geospatial data formats (e.g., Shapefiles, GeoJSON, GeoTIFF for MassGIS integration), tabular and statistical data (CSV, Parquet, HDF5, SAS, Stata), text and document formats (JSON, XML, PDF, unstructured text), time-series and sensor/IoT data (CSV, NetCDF, HDF5, streaming APIs), clinical and biomedical standards (HL7, FHIR, OMOP CDM, genomic VCF/FASTQ), and multimedia formats (JPEG, PNG, WAV, MP3, video containers like AVI/MP4). Vendors are encouraged to recommend a flexible and extensible approach that ensures interoperability with commonly used standards across healthcare, transportation, climate, geospatial, and other priority domains.

Question 205: Are there specific user concurrency requirements for different user types (researchers, agency staff, external partners) that would impact infrastructure sizing?

Answer: No, concurrency requirements have not been researched and defined. Project scope should include recommending a solution in partnership with stakeholders.

Question 206: How many datasets have you targeted to have at the end of year 1? What is the estimated size of the total data sets?

Answer: The DCC Requirements document lists potential datasets and provides illustrative size estimates. A specific target for the number and total size of datasets at the end of year 1 has not been finalized; part of the project scope is to define the initial data strategy and recommend an appropriate approach for onboarding and scaling datasets.

Question 207: What is the expected cadence and decision-making authority of the Data Access & Use Committee in reviewing dataset submissions (e.g., monthly vs. ad hoc)?"

Answer: The Data Access and Use Committee is detailing an intake and prioritization framework for dataset submissions.

Question 208: Will open access to datasets be allowed for non-sensitive data, or will all data require controlled or restricted access processes?

Answer: Open access to non-sensitive datasets may be allowed.

Question 209: What level of data cleaning, anonymization, or harmonization is expected from the consultant versus being left to dataset contributors?

Answer: The specific division of responsibilities for data cleaning, anonymization, and harmonization between the consultant, data owners, and users has not yet been fully decided. It is anticipated that vendors will propose approaches, which may include options such as dataset contributors providing these services themselves or data users complemented by consultant support to ensure consistency, quality, and compliance across the platform.

Question 210: Could you provide more information on the types of biases (e.g., algorithmic, data, or societal) that are of most concern to the Massachusetts AI Hub? Are there any specific, pre-approved AI fairness tools or frameworks that should be used or considered?

Answer: The Massachusetts AI Hub is concerned with a broad range of potential biases, including data bias (e.g., incomplete or unrepresentative datasets), algorithmic bias (e.g., model design and training disparities), and societal bias (e.g., inequitable impacts across demographic or socioeconomic groups). No specific AI fairness tools or frameworks have been pre-approved; vendors are encouraged to recommend and justify tools, methodologies, and best practices that can effectively detect, mitigate, and report on these biases in alignment with the Hub's goals for equity and transparency.

Question 211: Are there expectations as to the transfer rates of the data set both upload and download?

Answer: Specific transfer rate expectations for dataset upload and download have not yet been defined. Vendors should propose performance targets and scalable approaches that ensure efficient, reliable, and secure data transfer consistent with best practices and user needs.

Question 212: Should the platform support event-driven data pipelines (Kafka/Event Hub) or just batch ingestion?

Answer: Specific ingestion modes have not been mandated. The platform should be designed to support both batch and event-driven pipelines (e.g., Kafka/Event Hubs) where appropriate; vendors are encouraged to recommend an MVP approach (likely batch-first) with a roadmap for streaming/event-driven capabilities based on use-case needs, scalability, and cost.

Question 213: Which cloud region/data residency rules apply (must data be stored in Massachusetts)?

Answer: By default, all data (including logs, backups, and model artifacts) will be stored and processed in U.S. regions with an East-coast preference and the option to pin workloads to

MGHPCC (on-prem/private) for sensitive use cases or as appropriate in cloud service. As it pertains specifically to Commonwealth data, Contractor shall not allow its personnel or subcontractors to store Commonwealth Data on portable devices, including personal computers, except for devices that are used and kept only at its data centers located in the continental United States. Remote access to Commonwealth Data from outside the continental United States, including remote access to Commonwealth Data by authorized Services support staff in identified support centers, is prohibited unless approved in advance in writing by the Commonwealth.

Question 214: Should the system support schema-on-read (flexible data lake) or schema-on-write (strict data warehouse)?

Answer: This requirement has not yet been defined. Vendors should recommend an approach—schema-on-read, schema-on-write, or a hybrid—based on anticipated needs, governance, performance, and cost.

Question 215: Should the ingestion pipelines include automated data profiling (missing values, skew detection, outlier reporting)?

Answer: Not mandated, but recommended

Question 216: Should the pipelines include data quality scoring dashboards?

Answer: Not mandated, but recommended

Question 217: Should ingestion support streaming datasets (real-time feeds like MBTA data) or only static files?

Answer: Real-time feeds are out of scope for this phase of work. This functionality is planned for future releases (e.g., year two or three)

Question 218: Which data types must be supported first — tabular, text, image, video, time-series?

Answer: For the first release, the Data Commons is focused on foundational data types that cover common, high-impact use cases.

Question 219: Will there be a centralized ethics committee for dataset/model approvals, or distributed governance across data partners?

Answer: Yes, these are established committees.

Question 220: Should federated data access (models train on external data without moving it) be supported in Phase 1 or only later?

Answer: Federated data access is in scope for this project. See DCC Requirements Section 12

Question 221: Could you clarify the types of reports and dashboards you require?

Answer: Specific dashboard and reporting requirements have not yet been finalized. Vendors should propose a flexible reporting framework and include scope within the project to define detailed reporting needs in collaboration with MassTech. At a minimum, the platform should provide fundamental reports such as dataset usage and access metrics, system performance and uptime statistics, user activity and growth trends, and Commons Credits consumption summaries, with the ability to add more specialized dashboards as requirements evolve.

Question 222: Could you please specify different types of users who will interact with the platform? Additionally, we would like to understand the user journeys for each user type, particularly workflows, approval processes, and data visibility are expected to vary across these roles.

Answer: Below are different user profiles and required capabilities. User journeys and workflows have not yet been defined.

1. Researchers (Academic/Industry)
Responsibilities: Discover, access, and analyze datasets; contribute metadata; share results in compliance with terms of use.
Permissions:
* Search/browse full catalog.
* Access public and partner-approved datasets (tiered by license).
* Contribute datasets, annotations, and quality feedback.
* Request compute resources and synthetic data generation.

2. Industry Partners / Startups
Responsibilities: Use DCC data for innovation and commercialization; contribute private datasets or synthetic data pilots.

Permissions:
* Access datasets based on licensing tier (public, partner, restricted).
* Participate in pilot projects and Commons Credits incentives.
* Contribute datasets subject to approval workflows.

3. Educators & Students
Responsibilities: Use the DCC for learning, curriculum development, and research projects.
Permissions:
* Access open/public datasets.
* Leverage tutorials, synthetic data, and sandboxed environments.

4. Administrators (MA AI Hub / Delegated Stewards)
Responsibilities: Governance, policy enforcement, and system management.
Permissions:
* Approve/reject datasets, manage metadata quality, and enforce licensing.
* Configure Commons Credits rules, quotas, and access tiers.
* View full system audit trails and compliance reports.
* Manage delegated administrators at institutions or agencies.

5. Public / General Users
Responsibilities: Browse and discover publicly available datasets; provide feedback.
Permissions:
* Access and download open/public datasets.
* View metadata and documentation.

Question 223: Please clarify the expected level of dataset explorer functionality — should previews support only tabular/text data (CSV/JSON) or also images, geospatial, and video?

Answer: The required level of dataset explorer functionality has not yet been finalized. Vendors should propose an approach that, at a minimum, supports previews of tabular and text data (e.g., CSV, JSON) and can be extended to include images, geospatial data, and video as the platform evolves and user needs expand.

Question 224: What security measures are currently in place? Are there any specific compliance requirements for the new portal that need to be met?

Answer: No existing security measures are in place, vendors should recommend them as part of the engagement.

Question 225: What is the required frequency for data backups (e.g., real-time, daily, weekly or monthly)?

Answer: No requirement is currently in place, vendors are invited to recommend approaches that follow best practices and support the goals the DCC

Question 226: What is the expected data retention policy (e.g., how long should backups be stored)?

Answer: This has not been defined. Vendors should recommend policies based on best practices.

Question 227: Please specify the SLA for maintenance and support.

Answer: Ongoing 24/7 system support is not in scope for this RFP. Vendors are expected to provide support during the design, build, testing, and initial rollout phases, including documentation, training, and assistance with launch. Future operational support models

Question 228: Data Provenance & Lineage: The requirement for "provenance tagging" is noted. What is the desired depth of provenance? Is it sufficient to track the origin and basic transformations of a dataset, or is a full, granular data lineage capability required?

Answer: The desired depth of provenance and lineage has not been predefined. Vendors should recommend the appropriate level of tracking—ranging from dataset origin and basic transformations to more granular lineage—based on best practices, use cases, and governance needs.

Question 229: Collaboration with Other Vendors: Will there be other vendors involved in this project, and how will collaboration be managed among different parties?

Answer: We prefer to award this contract to a single with sub-contractors (if needed). Other stakeholders are involved including Massachusetts state agencies, the MGHPCC and AICR consortium institutions, and the broader Massachusetts innovation ecosystem. Respondents should anticipate engaging across this multi-sector community to ensure that the platform reflects diverse needs and encourages wide adoption. The MA AI Hub will facilitate introductions to these stakeholders and coordinate engagement as appropriate.

Operations & Monitoring

Question 230: There is a deliverable titled “Pilot user feedback,” what is the pilot referenced in this deliverable?

Answer: Collecting early user feedback in a pilot to understand their experiences, challenges, and suggestions is in scope. This feedback will be analyzed to refine the system, improve usability, and guide decisions for broader deployment.

Question 231: Should the site support offline or low-bandwidth modes for mobile users?

Answer: Not in scope for this phase.

Question 232: Can MassTech please clarify which "suggested initial datasets" from Section 9.2 are mandatory vs optional?

Answer: The datasets referenced in Section 9.2 are suggested examples to illustrate the types of data that may be valuable for the initial implementation of the Data Commons. They are not mandatory; vendors may propose alternative or additional datasets they believe would best demonstrate the platform’s capabilities and serve priority use cases.

Question 233: Are any tools approved/forbidden for initial configuration (e.g., SDV, Synthea, CARLA, Gretel, YData, Ground Truth Synthetic)?

Answer: No tools have been pre-approved or forbidden for initial configuration. The tools listed in the RFP are illustrative examples, and vendors are encouraged to recommend the synthetic data tools and frameworks they believe are most appropriate to meet the goals of the Data Commons.

Question 234: Which specific Massachusetts agencies are expected to be initial participants in the DCC, and what are their current data sharing agreements and technical capabilities?

Answer: The MA AI Hub coordinates the MA agency participants in the Data Commons Collaborative.

Question 235: Are there specific data classification schemes or sensitivity levels already in use across Massachusetts agencies that the DCC must adopt or accommodate?

Answer: Yes. The DCC will adhere to data classification security standards as set forward by the MA CISO.

Question 236: What is the expected data volume at launch and projected growth over 3 years? This impacts infrastructure sizing and cost modeling for Snowflake and computing resources.

Answer: The expected data volume at launch and projected growth over three years have not yet been defined. Vendors should propose assumptions, scenarios, and scalable infrastructure approaches to support initial operations while accommodating future growth in alignment with best practices and cost-effective resource planning.

Question 237: What specific protected attributes and fairness metrics are prioritized for Massachusetts use cases (beyond race, gender, age mentioned in Section 12.1), particularly for state-specific demographics and socioeconomic factors?

Answer: Specific protected attributes and fairness metrics beyond those noted in Section 12.1 have not yet been finalized. However, it is anticipated that in addition to race, gender, and age, Massachusetts use cases may require consideration of factors such as ethnicity, disability status, socioeconomic, education, geographic location (urban/rural), and language access, to ensure equitable outcomes across diverse state populations. Vendors are encouraged to recommend fairness metrics and methodologies that account for these demographics and socioeconomic factors, while aligning with best practices and applicable regulatory guidance.

Question 238: What is the expected budget range for ongoing operations post-development, and what cost-sharing model is envisioned between Massachusetts agencies and external users?

Answer: Operations of the data commons will be supported through AICR and some level of cloud compute as needed. A cost-sharing model will be examined as needs scale.

Question 239: Are there specific requirements for financial transparency and cost breakdowns that must be provided to participating agencies for budget planning purposes?

Answer: We encourage vendors to submit competitive proposals with clear and transparent pricing, including breakdowns of one-time and ongoing costs as well as any optional features.

Question 240: Are there specific metadata standards or ontologies already in use by Massachusetts agencies that must be incorporated into the data catalog?

Answer: No specific metadata standards are mandated at launch. Vendors should recommend appropriate standards (e.g., Dublin Core, DCAT, Schema.org, or others) and propose a practical minimum completeness bar that balances usability, interoperability, and ease of implementation for the initial launch.

Question 241: Are there specific future capabilities or integrations already identified in the Massachusetts AI Hub strategic roadmap that should influence the initial architecture design?

Answer: Yes. Though not envisioned for inclusion in the scope of this development and implementation, an expanded functionality to include prompt and model usage tracking (e.g., logging how models are invoked and linking outputs back to specific runs, with appropriate privacy safeguards) as well as model usage dashboards that give administrators visibility into dataset utilization, model calls, and overall resource consumption is a desired future state. Designing for easy integration of these capabilities into the platform is desired. The design should allow the possible extension of the Commons Credits incentive system to potentially include sustainability metrics and commercialization .

Question 242: Are there any existing user stores that need to be integrated into the AI Hub? If yes, please provide a summary of user repositories to be integrated.

Answer: No existing user stores are in place

Question 243: Can you provide an estimate of the expected training audience size (e.g., number of administrators, researchers, public users)?

Answer: An estimate of the expected training audience size, including numbers of administrators, researchers, and public users, has not yet been defined. Determining training group sizes, along with the overall training approach and plan, will be part of the project scope, and vendors are encouraged to recommend best-practice strategies to address diverse user needs.

Question 244: What is the expected user volume over the first year and what is the target going into the future?

Answer: The RFP and DCC Requirements do not prescribe exact user counts, but the MA AI Hub anticipates a broad and steadily growing user base. In the first year, the platform is expected to onboard hundreds of users growing to thousands of users across Massachusetts agencies, researchers, and academic partners, with additional adoption from startups and industry as awareness builds.

Question 245: Will the initial dataset list provided in the DCC requirements be considered mandatory for inclusion, or is it illustrative only?

Answer: The data set listed in the DCC Requirements document is illustrative

Question 246: Are there state cybersecurity standards we must align with beyond general encryption, MFA, and RBAC?

Answer: Massachusetts-specific cybersecurity standards beyond general practices such as encryption, MFA, and RBAC have not yet been defined for the DCC. Vendors should align with recognized best practices and frameworks and be prepared to incorporate any state-specific requirements identified during the course of the project.

Question 247: What are the specific data sources and formats that the Massachusetts AI Hub intends to use for the initial datasets? Will there be a complete list of these datasets provided and access to them prior to the project's start, to inform the proposed architecture?

Answer: The DCC Requirements document contains a list of potential data sources. In scope for the consulting engagement is finalizing the data sources. The proposed solution should be designed to support a wide variety of data sources and dataset types and formats, including but not limited to tabular data (CSV, Parquet), geospatial data (Shapefiles, GeoJSON, GeoTIFF for MassGIS integration), healthcare and biomedical data (HL7, FHIR, DICOM, NIfTI, genomic VCF/FASTQ), transportation and IoT/time-series data (JSON, NetCDF, streaming APIs), and multimedia data (MP4, JPEG, PNG, WAV). Vendors are encouraged to recommend architectures that ensure flexibility, scalability, and interoperability to accommodate these diverse data sources.

Question 248: Will MassTech provide pilot users for feedback, or is the respondent expected to recruit and manage pilot participants?

Answer: MassTech will identify and recruit pilot users from the stakeholder community.

Question 249: Can you clarify what is expected in the "Explanation of planned industry engagement"? Is this intended as a marketing plan, a partnership strategy, or an outreach roadmap?

Answer: The required deliverable, "Explanation of planned industry engagement" is the expectation to include a narrative describing how industry stakeholders (startups, private sector partners, research organizations) will be engaged in the platform's development and adoption process. It is not marketing plan; rather a strategic engagement roadmap detailing how the bidder plans to involve, inform, and collaborate with industry customers of the DCC to drive relevance, contribution, and adoption.

Question 250: Should training materials be designed for technical users only, or will there also be a need for non-technical stakeholder training?

Answer: Training materials should be designed to support both technical users and non-technical stakeholders. Both groups may be in scope, and the final training approach will be defined as part of the project to ensure accessibility and usability across the full range of anticipated users.

Question 251: What are the expected API throughput and concurrency requirements (e.g., 100 req/sec, 1000+ concurrent users)?

Answer: The RFP and DCC Requirements do not prescribe exact user counts, but the MA AI Hub anticipates a broad and steadily growing user base. In the first year, the platform is expected to onboard hundreds of users growing to thousands of users across Massachusetts agencies, researchers, and academic partners, with additional adoption from startups and industry as awareness builds.

Question 252: What is the expected approval workflow for publishing datasets (central review committee vs automated validation)?

Answer: Initially review committee.

Question 253: What is the expected scale of transactions/users in Phase 1 (hundreds vs thousands)?

Answer: Scaling from hundreds to thousands is expected in the first year.

Question 254: Should it integrate with identity providers (SSO, OAuth, eduGAIN for universities) for user-level tracking?

Answer: The proposed system should have the ability to confirm user identity to enable accurate, secure tracking of Commons Credits. However, no specific identity-management approach is mandated; vendors are encouraged to recommend solutions that best meet this goal.

Question 255: Should fairness reporting integrate with external regulators or compliance agencies?

Answer: This requirement has not yet been defined. For the purposes of the RFP, vendors should submit proposals that fully address the requested scope and design solutions with the flexibility to integrate with external partners if needed.

Question 256: Which specific Massachusetts/state cybersecurity frameworks must be adhered to (beyond RBAC, MFA)?

Answer: Specific Massachusetts or state-level cybersecurity frameworks beyond RBAC and MFA have not yet been defined for the DCC. Vendors should align with recognized best practices and standards (e.g., NIST, CISA) and be prepared to incorporate any state-specific requirements identified during the project.

Question 257: Should governance align with NIST AI RMF, EU AI Act, Algorithmic Accountability Act, or only state/federal US laws?

Answer: Governance must comply with applicable state and U.S. federal laws and regulations. Vendors are encouraged to align their frameworks with widely recognized standards such as the NIST AI Risk Management Framework.

Question 258: Who are the initial pilot users (universities, startups, state agencies), and expected user volume at launch?

Answer: The initial users are anticipated to come from universities, hackathon participants, and / or startups. MassTech will determine pilot users. User volumes will scale from hundreds to thousands.

Question 259: Should the platform provide sandbox environments for universities/startups (isolated tenants)?

Answer: Yes.

Question 260: Should the system include usage analytics dashboards (dataset popularity, query performance, user activity)?

Answer: Yes, system usage reporting is required.

Question 261: Can you provide a detailed breakdown of user roles (such as researcher, developer, public agency staff) along with their specific responsibilities and permissions?

Answer: 1. Researchers (Academic/Industry)
Responsibilities: Discover, access, and analyze datasets; contribute metadata; share results in compliance with terms of use.
Permissions:
* Search/browse full catalog.
* Access public and partner-approved datasets (tiered by license).
* Contribute datasets, annotations, and quality feedback.
* Request compute resources and synthetic data generation.

2. Industry Partners / Startups
Responsibilities: Use DCC data for innovation and commercialization; contribute private datasets or synthetic data pilots.
Permissions:
* Access datasets based on licensing tier (public, partner, restricted).
* Participate in pilot projects and Commons Credits incentives.
* Contribute datasets subject to approval workflows.

3. Educators & Students
Responsibilities: Use the DCC for learning, curriculum development, and research projects.
Permissions:
* Access open/public datasets.
* Leverage tutorials, synthetic data, and sandboxed environments.

4. Administrators (MA AI Hub / Delegated Stewards)
Responsibilities: Governance, policy enforcement, and system management.
Permissions:
* Approve/reject datasets, manage metadata quality, and enforce licensing.
* Configure Commons Credits rules, quotas, and access tiers.
* View full system audit trails and compliance reports.
* Manage delegated administrators at institutions or agencies.

5. Public / General Users
Responsibilities: Browse and discover publicly available datasets; provide feedback.
Permissions:
* Access and download open/public datasets.
* View metadata and documentation.

Question 262: Should the architecture support multi-tenant isolation (e.g., separate research groups, universities) or a shared commons only?

Answer: Multi-tenant Administration

Question 263: What is the tentative user count for concurrent users, active users, and daily user activity for now and in the future?

Answer: The MA AI Hub anticipates a measured ramp-up in adoption. In the initial launch year, the platform should be designed to support hundreds of concurrent users. Looking ahead, as industry and regional adoption grows, the target is to scale to 1,000+ concurrent users. The architecture should therefore be elastic, able to expand capacity seamlessly as adoption accelerates.

Question 264: Pilot Users & Feedback: For Deliverable #10, "Pilot user feedback," will MassTech be responsible for identifying and recruiting pilot users from the stakeholder community, or is the selected respondent expected to manage that process?

Answer: MassTech will collaborate with the Vendor on pilot user outreach.

Question 265: Initial Dataset Curation: For "curating the initial set of datasets," what is the expected level of effort from the respondent? Will MassTech provide a list of

target datasets and sources, or is the respondent expected to proactively identify and propose high-value datasets for inclusion?

Answer: The respondent is expected to identify and propose datasets using provided illustrative lists and in partnership with MassTech

Question 266: Stakeholder Identification: Beyond the MassTech team, who are the key stakeholder groups (e.g., specific state agencies, university partners, industry groups) that the respondent will be expected to coordinate with, and will MassTech facilitate introductions?

Answer: Beyond the MA AI Hub/MassTech team, the key stakeholder groups for the DCC include: Massachusetts state agencies (e.g., health, transportation, education, and economic development entities that steward high-value datasets), the MGHPCC and AICR consortium institutions, and the broader Massachusetts innovation ecosystem, including startups, established industry partners, academic institutions and nonprofit organizations. Respondents should anticipate engaging across this multi-sector community to ensure that the platform reflects diverse needs and encourages wide adoption. The MA AI Hub will facilitate introductions to these stakeholders and coordinate engagement where appropriate, while also expecting the awarded vendor to take a proactive role in stakeholder management and collaboration.

Question 267: Presentation Format: If selected as a finalist, what is the expected format and duration of the presentation? Will it be virtual or in-person?

Answer: If MassTech chooses to have oral presentations, vendors would be expected to provide a PowerPoint presentation and / or demonstration

Question 268: Which accessibility standard should we certify against at launch (e.g., WCAG 2.1 AA), and under what user/concurrency assumptions should the <2s page/search targets be tested?

Answer: Yes, the platform should incorporate accessibility features that ensure equitable access across languages. At minimum, compliance with WCAG 2.1 accessibility standards will be required, including support for screen readers, captioning, and text alternatives. In addition, multilingual accessibility is an important consideration given the diverse communities expected to use the platform. This means that content, metadata, and user interfaces should be designed to accommodate non-English languages, and screen reader compatibility must extend beyond English text. While the initial scope may prioritize English, the architecture should be built to enable future localization and translation, ensuring that multilingual accessibility can be expanded as the user base grows.

Question 269: Is work expected to be performed remote or onsite ?

Answer: Work on the project may be performed either remotely or onsite. Vendors are encouraged to propose the delivery approach that best supports effective collaboration, communication, and successful execution of the project.

Question 270: Do you have a specific minimum or maximum on the quantity / data volume of initial datasets?

Answer: No. Data volume has not yet been finalized

Question 271: Are the primary users of these tools the DCC team or will they be open to all/subset of AI Hub users?

Answer: The DCC capabilities will be available to a breadth of users. See DCC Requirements Section 4.0 for a notional list of users.

Question 272: If tool use is open to external users outside of the DCC team/Commonwealth employees, will users have the ability to store "draft" synthetic datasets for their own use before going through a publishing process?

Answer: Yes, users should have the ability to generate and store synthetic datasets

Question 273: Are all users expected to be authenticated or will there be anonymous access permitted?

Answer: All users are expected to be authenticated

Question 274: Does the Commonwealth of Massachusetts state organization as well as other participating organizations have a preference regarding using a data catalog provider?

Answer: No preference currently exists, MassTech is open to recommendations

Question 275: Does the Commonwealth of Massachusetts state organization as well as other participating organizations have a preference for software as a Service and/or on-premise solution(s) of the technologies listed?

Answer: There is no current preference between Software-as-a-Service (SaaS), on-premise, or hybrid deployment models for the technologies listed. Vendors should propose the approach, cloud-based, on-premise, or a combination, that best meets the project's goals for scalability, security, cost-effectiveness, and long-term sustainability.

Project Management & Staffing

Question 276: What is Mass Tech's current tech stack/tech preferences?

Answer: No preferences (MA AI Hub).

Question 277: What licenses does Mass Tech's currently hold?

Answer: None (MA AI Hub).

Question 278: Does Mass Tech have an anticipated completion date for this project?

Answer: MassTech does not have a fixed completion date for this project. Vendors should propose a realistic timeline with clear milestones that reflect the scope and complexity of their solution.

Question 279: Does Mass. Tech anticipate that the solution will include data lifecycle policies, so that data sets might age out or expire over time?

Answer: Yes, the solution is expected to include data lifecycle policies to manage datasets over time. This should address processes such as data aging, expiration, archival, or deletion, in alignment with governance, compliance, and user needs.

Question 280: Does Mass. Tech anticipate that the data synthesis requirements will be covered by the tools listed or new tooling will need to be developed to satisfy needs?

Answer: The tools listed in the RFP are provided as illustrative examples and may address some requirements, but they are not prescriptive. MassTech anticipates that vendors may propose the use of existing tools, new tooling, or a combination of both, based on what best satisfies the functional, scalability, and compliance needs of the Data Commons.

Question 281: How does Mass. Tech define the impact of bias?

Answer: For the MA AI Hub, the impact of bias is defined in terms of the downstream harms or inequities that arise when datasets or AI models reflect or amplify unfair patterns. This includes effects on accuracy, fairness, and representation across demographic groups, as well as risks of systematic exclusion, discrimination, or misallocation of resources. In the context of the Data Commons Collaborative, assessing bias impact means evaluating not just whether bias exists in data or models, but also how that bias could influence research outcomes, policy decisions, or commercial applications.

Question 282: What is the expected duration of the engagement? Is there a target start and end date?

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 283: Could you please confirm what is the target go-live date or public launch timeline for the Data Commons Collaborative platform?

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 284: Contract Term & Value: What is the anticipated period of performance for this consulting engagement? Is there an anticipated start date and a target date for the "public launch" deliverable?

Answer: The anticipated start date for the engagement is December 2, 2025. The MA AI Hub is not prescribing specific phase durations, as we recognize that different vendors may propose varying approaches to achieve the deliverables outlined in the RFP and the DCC Requirements. A phased release is preferred in order to get the Data Commons Collaborative (DCC) live and usable sooner, while ensuring stability and stakeholder engagement. Vendors are encouraged to include recommendations for a phased release roadmap in their proposals, showing how to balance early usability with long-term extensibility.

Question 285: Are you open to a single proposal from a primary organization with one or more named partners (e.g., consultant + a tech partner)?

Answer: Yes.

Question 286: Should Attachment C include infrastructure/compute/licenses (e.g., cloud/GPU, catalog or fairness tool licenses), or is pricing services-only? Beyond mileage, are any other reimbursables allowable?

Answer: The Attachment C - Budget Template should include all costs that are part of the proposed project. The budget is not limited to services-only. All reimbursable expenses should be described in the proposal so that MassTech can evaluate their appropriateness.

Question 287: Are you looking to award the engagement to 1 entity with cross-industry/sectors expertise, OR are you exploring creating a group of industry/sector experts with shared expertise in AI/Data/tech enablement to serve you?

Answer: We prefer to award this contract to a single with sub-contractors (if needed)

Question 288: Are there preferences or constraints regarding open-source vs. proprietary components (e.g., Dataverse vs. commercial metadata tools)?

Answer: There are no specific preferences or constraints regarding the use of open-source versus proprietary components. Vendors should recommend the mix of technologies—open-source, commercial, or hybrid—that best meets the Data Commons goals for scalability, interoperability, sustainability, and cost-effectiveness.

Question 289: Are the examples of tools/technologies listed in the solicitation currently part of or expected to be part of the future tech stack?

Answer: The RFP and DCC Requirements list examples of tools and technologies to illustrate the types of capabilities expected, but they are not mandated components of the technology stack. These examples should be treated as illustrative references, showing the kinds of functions the platform must support.

Security & Compliance

Question 290: Can MassTech please clarify if there are specific compliance frameworks that need to be supported beyond HIPAA?

Answer: There are not currently identified compliance frameworks, vendors are invited to recommend requirements based on project objectives and goals.

Question 291: Reproducible Research - Because the platform is expected to support 'ethical, fair, and secure AI development,' will it also help researchers keep track of their experiments and results—like using model cards, data cards, or experiment tracking tools (for example, MLflow, Weights & Biases, or Hugging Face Hub)?

Answer: Because the platform is designed to support ethical, fair, and secure AI development, reproducibility is a priority from the outset. We plan to adopt a hybrid approach that balances early usability with long-term capability.

At launch, the Data Commons should include a baseline metadata layer that captures essential reproducibility information such as lightweight model cards, data cards, and persistent experiment identifiers. This ensures that researchers can document key assumptions, methods, and outputs in a consistent and transparent way from the start.

Over time, we envision expanding functionality by enabling integration with widely used experiment-tracking tools such as MLflow, Weights & Biases, and Hugging Face Hub. These integrations will allow researchers to maintain their preferred workflows while still linking results into the Data Commons for broader accessibility, comparability, and trust.

This phased approach ensures that reproducibility is embedded as a guiding principle from day one, while leaving room for the ecosystem to grow into more advanced experiment-tracking services as adoption and needs evolve

Question 292: AI Fairness and Bias - Because the platform will 'implement bias detection, auditing, and mitigation tools' and promote 'fair and secure AI development,' will users be able to test how fair their AI models are—like checking for bias in outputs or datasets (for example, using AIF360, Fairlearn, or prompt-level toxicity scoring for LLMs)?

Answer: Yes, the system should have tools for ensuring AI fairness using tools like bias detection, monitoring, etc. See DCC Requirements Section 11

Question 293: Security, Governance & Compliance - Given the platform will 'define and implement RBAC, MFA, and dataset tiering' and enforce 'ethical data handling,' will it also track how AI models and prompts are used—like logging who accessed what data or which model generated which output (for example, using audit trails, prompt logging, or model usage dashboards)?

Answer: Yes. In addition to implementing RBAC, MFA, and dataset tiering, the platform should support baseline audit and usage tracking from the outset. This includes audit trails that log who accessed which datasets, when, and under what authorization. These foundational capabilities ensure ethical data handling and compliance are embedded into the system.

Though not envisioned for inclusion in the scope of this development and implementation, an expanded functionality to include prompt and model usage tracking (e.g., logging how models are invoked and linking outputs back to specific runs, with appropriate privacy safeguards) as well as model usage dashboards that give administrators visibility into dataset utilization, model calls, and overall resource consumption is a desired future state. Designing for easy integration of these capabilities into the platform is desired.

Question 294: Compute Requirements - Because the platform will support tools like 'SDV, Synthea, CARLA' and potentially bias detection frameworks, do you expect it to require dedicated compute resources—like CPU or GPU clusters for data processing and model execution (for example, using high-performance servers with AI accelerators or virtualized compute nodes)?

Answer: A hybrid setup of GPUs and cloud computing will support the data commons in addition to end user compute platforms.

Question 295: What metadata standards are required? Any custom fields needed for cataloging?

Answer: Meta data needs are highlighted in DCC Requirements Section 8. The project scope includes recommending and implementing standards.

Question 296: Should metadata enrichment include automated tagging, quality scoring, or bias flagging?

Answer: For launch, metadata enrichment should include automated tagging and dataset quality scoring. The framework should also leave room for bias and fairness metadata, which can be integrated as the fairness dashboards and auditing tools described in the RFP mature.

Question 297: Does the workflow require multi-tenant or delegated catalog administration?

Answer: Multi-tenant Administration
Each participating institution, program, or partner may need to curate, publish, and manage its own datasets while maintaining clear separation from other tenants. Multi-tenant administration ensures role-based access to catalogs, metadata, and submission workflows while preserving isolation of sensitive datasets.

Delegated Catalog Administration
Beyond a central administrator, the platform should support delegated roles (e.g., dataset stewards, institutional leads, project owners) who can review, approve, or enrich submissions within their domain. This matches the requirement for an Admin Dashboard with submission management and workflow oversight.

Question 298: Are there any fairness/bias detection frameworks that should be prioritized for integration? e,g, AIF360, Fairlearn

Answer: Project scope includes recommending, designing, and implementing fairness/bias detection capabilities. Anticipated in this scope are frameworks for data set auditing, bias mitigation, model explainability, drift detection, etc. Frameworks such as AIF360, Fairlearn and others can be included (e.g., Jurity, SHAP, etc.)

Question 299: Are there requirements for live bias detection in data pipelines or only during audits?

Answer: For the scope of this effort, audit-time checks now, with architecture hooks for live bias detection desired so it can be turned on per-pipeline without a redesign

Question 300: Are the evaluation criteria stated in "4.2 Criteria" listed in descending order of importance, ascending order of importance, or are they equally weighted?

Answer: The criteria identified are provided as a comprehensive set of evaluation factors and are not intended to represent a prioritized or weighted list. The Commonwealth will apply these criteria in a balanced and holistic manner to assess the qualifications, capabilities, and overall value of each submitter in determining the best interest of the Data Commons initiative.

Question 301: Is there a budget for this project?

Answer: Funding has been allocated for this project, but a specific budget ceiling is not being disclosed at this stage. We encourage vendors to submit competitive proposals with clear and transparent pricing, including breakdowns of one-time and ongoing costs as well as any optional features.

Question 302: Are there specific devices or screen sizes that must be prioritized for testing?

Answer: No.

Question 303: What specific accessibility standards must be met (e.g., WCAG 2.1 AA)?

Answer: The MA AI Hub expects the Data Commons Collaborative to comply with recognized digital accessibility standards, including WCAG 2.1 Level AA at a minimum, consistent with Commonwealth of Massachusetts IT accessibility policies. This means the platform must be usable by individuals with disabilities across key areas such as vision, hearing, motor, and cognitive access. In addition, all web-based services must support screen readers, keyboard navigation, and captioning/transcripts for multimedia content. Vendors are encouraged to design with universal accessibility principles so that the system not only meets compliance benchmarks but also provides an inclusive experience for all users.

Question 304: Will there be a third-party accessibility audit or certification required?

Answer: The RFP and Requirements do not mandate a formal third-party accessibility certification. However, the MA AI Hub expects vendors to demonstrate compliance with WCAG 2.1 AA and Commonwealth accessibility policies through their development and QA processes. Vendors should plan to conduct internal accessibility testing (including assistive technology compatibility and user acceptance testing with accessibility in mind) and provide documentation of results. The MA AI Hub may, at its discretion, engage an independent accessibility review as part of acceptance testing or future audits, but this is not specified as a baseline requirement in the current RFP.

Question 305: What are the specific audit and reporting requirements for the Ethics Oversight Board mentioned in Section 12.3, and how frequently must bias impact assessments be conducted and reported?

Answer: While the ultimate responsibility for executing these requirements will rest with the system operator, vendors should demonstrate in their proposals how their approach, tools, or governance models would enable and support recurring audit and reporting obligations.

Question 306: Are there specific training and certification requirements for users accessing the DCC, particularly for ethical AI practices and bias detection?

Answer: Yes, training and certification requirements for users—particularly in areas such as ethical AI practices, responsible data use, and bias detection—are expected to be required. However, the specific scope, structure, and standards for these requirements have not yet been defined and will be developed as part of the project.

Question 307: What are the specific cybersecurity frameworks or standards (e.g., NIST, CISA) that must be implemented, and are there Massachusetts-specific security assessment requirements?

Answer: Specific cybersecurity frameworks and standards for the DCC have not yet been mandated. Vendors should expect to align with widely recognized frameworks such as NIST and CISA guidance, while also meeting any applicable Massachusetts state security assessment requirements as they are defined during the project. Vendors are encouraged to recommend best-practice approaches that ensure compliance, resiliency, and protection of sensitive data.

Question 308: What are your accessibility requirements levels? (e.g., Legal and compliance standards, content accessibility, etc.)

Answer: The MA AI Hub expects the Data Commons Collaborative to comply with recognized digital accessibility standards, including WCAG 2.1 Level AA at a minimum, consistent with Commonwealth of Massachusetts IT accessibility policies. This means the platform must be usable by individuals with disabilities across key areas such as vision, hearing, motor, and cognitive access. In addition, all web-based services must support screen readers, keyboard navigation, and captioning/transcripts for multimedia content. Vendors are encouraged to design with universal accessibility principles so that the system not only meets compliance benchmarks but also provides an inclusive experience for all users.

Question 309: What languages will you be supporting?

Answer: The platform should incorporate accessibility features that ensure equitable access across languages. At minimum, compliance with WCAG 2.1 accessibility standards will be required, including support for screen readers, captioning, and text alternatives. In addition, multilingual accessibility is an important consideration given the diverse communities expected to use the platform. This means that content, metadata, and user interfaces should be designed to accommodate non-English languages, and screen reader compatibility must extend beyond English text. While the initial scope may prioritize English, the architecture should be built to enable future localization and translation, ensuring that multilingual accessibility can be expanded as the user base grows.

Question 310: Do you require any archival capabilities for logs after 3 years?

Answer: An archival strategy for logs beyond three years has not yet been defined. Vendors should consider archival capabilities as part of the project scope and propose recommended approaches based on best practices and compliance requirements.

Question 311: Have you identified the specific events and details that will be captured in the logs? This will be a driver of the overall cost associated with event logging and reporting.

Answer: The specific events and level of detail to be captured in the logs have not yet been defined. This will be determined as part of the project scope, and vendors should propose recommended approaches that balance functionality, compliance, and cost efficiency.

Question 312: Do you anticipate any of the sources be streaming data or should we assume batch?

Answer: Real-time feeds are out of scope for this phase of work. This functionality is planned for future releases.

Question 313: Do you have any defined guidelines or criteria on when you use open-source vs. commercial solutions?

Answer: No formal guidelines or criteria have been established to determine when open-source versus commercial solutions should be used. Vendors are encouraged to recommend approaches and tools—whether open-source or commercial—that best balance cost, scalability, security, supportability, and alignment with the goals of the Data Commons.

Question 314: What outcomes constitute success (e.g., number/type of partners, MOUs, events)?

Answer: Success for the DCC will be measured by both ecosystem engagement and user experience. Key outcomes include securing diverse partners across academia, government, startups, and industry; formalizing collaborations through MOUs; and convening events that drive adoption and data contributions. Performance metrics will focus on user satisfaction and usability survey results as well as measurable collaboration indicators such as the number of data exchanges between DCC users and the diversity of domains represented in those exchanges.

Question 315: Are there any penalties, liquidated damages, or termination clauses tied to project milestones or deliverables?

Answer: Please review MassTech's procurement webpage for examples of contract language used in various agreement templates. Specific language will be discussed as part of the contracting phase.

Question 316: Should fairness reporting be public-facing dashboards, or primarily for internal auditing and compliance?

Answer: For this phase of work, fairness reporting is intended primarily for internal auditing and compliance purposes. Public-facing dashboards are not required at this stage.

Question 317: There are a number of requirements in the solicitation that relate to defining policies and standards, for instance around ethical usage, bias detection and data tier delineation. Does the Mass. Tech Collaborative anticipate that defining policies like those will be a collaborative or consultative effort with the winning party, or is the Tech Collaborative looking for the winning party to lead on those questions more independently.

Answer: Defining policies and standards such as ethical usage, bias detection, and data tier delineation will be a collaborative effort.

Question 318: Should the budget be broken down by the eight service areas listed in Section 2, or by project phases (e.g., design, development, rollout)?

Answer: The budget should be completed using the provided template; it does not require a breakdown by the eight service areas or by project phases. However, Respondents may provide additional detail on budget by phase, milestone or service area as part of their project summary or narrative if they believe it will help clarify their approach.

Question 319: Is there an expectation of high availability (99.9%+ uptime) and disaster recovery (RTO/RPO targets)?

Answer: The system is not expected to operate as a full production-strength solution with strict high-availability requirements. Instead, it is intended to support research, experimentation, and initial AI deployments, with vendors encouraged to propose reasonable availability and disaster recovery approaches appropriate to this context.

Question 320: Should metadata records be versioned (history of schema/description/tags)?

Answer: Yes

Question 321: Is integration with graph databases (e.g., Neo4j) for dataset relationship mapping required?

Answer: Graph database integration has not been mandated. Vendors are welcome to recommend solutions to expand capabilities and functionality.

Question 322: Are there expectations around differential privacy guarantees (e-differential privacy budgets)?

Answer: The RFP does not mandate specific differential privacy guarantees or prescribe ε-differential privacy budgets. Vendors are expected to propose appropriate privacy protections for synthetic data generation, which may include differential privacy or comparable methods. Proposals should describe how privacy-utility tradeoffs will be managed and how contributors will be informed about the level of protection applied.

Question 323: What level of compliance certification is required (HIPAA, GDPR, CCPA)?

Answer: The RFP does not require vendors to hold formal HIPAA, GDPR, or CCPA certifications. However, the platform must be designed to support compliance with these and other applicable regulatory frameworks when handling sensitive data. Vendors should describe how their solution addresses relevant security and privacy requirements and how compliance can be demonstrated for datasets subject to specific regulations.

Question 324: Which fairness metrics are considered mandatory (e.g., disparate impact, equal opportunity, demographic parity)?

Answer: The expectation is that bias and fairness will be evaluated using recognized, widely accepted measures that are appropriate to the dataset, model, and use case. At a minimum, vendors should demonstrate familiarity with these and related fairness measures, and propose a framework that is transparent, repeatable, and defensible. Vendors are encouraged to justify their selection of metrics based on relevance to the application and to describe how multiple fairness perspectives will be incorporated into bias impact assessments.

Question 325: Should bias detection support subgroup analysis across multiple protected attributes simultaneously?

Answer: Vendors should propose bias detection approaches that effectively address the needs. The selected vendor is expected to act as the expert in recommending methods, tools, and metrics that best meet these requirements and ensure comprehensive fairness evaluation.

Question 326: Should fairness dashboards be designed for internal expert review or for public transparency?

Answer: Fairness dashboards should support both internal expert review and external transparency. Vendors should design dashboards that provide detailed metrics and analysis for governance and technical users, while also enabling public-facing transparency features (e.g., high-level fairness indicators or audit summaries) to promote trust without disclosing sensitive data.

Question 327: Should dashboards include interactive scenario testing (e.g., 'What if this group had 10% more samples')?

Answer: Not in scope for this phase.

Question 328: Is there a requirement for continuous monitoring of fairness drift in deployed models?

Answer: Yes

Question 329: Should fairness results be exportable in standard audit formats (PDF, JSON reports, regulatory templates)?

Answer: Yes

Question 330: How will fairness assessments be validated and approved (Fairness Oversight Board, technical auditors)?

Answer: An industry governance committee is in place and working to define an oversite approach. It is anticipated that the selected vendor will work with the committee to establish and implement fairness solutions.

Question 331: Should datasets be encrypted with separate keys per dataset/tenant?

Answer: This requirement has not yet been defined. Vendors should recommend an encryption strategy based on best practices for security, scalability, and compliance.

Question 332: For the AI fairness and bias framework, does MassTech expect real-time monitoring of deployed models, or only offline evaluation of datasets and research models?

Answer: Real-time monitoring is not required, but the system should not exclude future capability enhancements that include this functionality.

Question 333: How many staff members will require training? Are you fine with online training for the staff?

Answer: Specific training counts have not been defined. Online training is acceptable.

Question 334: Governance Model: The RFP tasks the respondent with defining governance and terms of use. What is the envisioned governance structure after launch? (e.g., Will there be a steering committee? Who will have the authority to approve access to restricted data tiers?).

Answer: A governance committee is already in place and will participate actively in the project. Beyond academic institutions, the governance framework is expected to include representatives from government, industry, startups, and community stakeholders to ensure broad input and transparency.

Question 335: Key Personnel: Are there specific key personnel roles (e.g., Program Manager, Lead Architect, Data Governance Specialist) that must be dedicated full-time to this engagement, or is a blended, fractional allocation model acceptable?

Answer: Personnel and staffing approaches are not mandated. Vendors should propose the key personnel roles, level of effort, and staffing model—whether full-time or blended/fractional—that they believe will best accomplish the requested scope and ensure successful delivery of the engagement.

Question 336: Are there any datasets we would create or cleanse as a part of this initiative, or will they all be sourced externally?

Answer: Datasets will be sourced externally (e.g., public data sets and Massachusetts state government datasets)

Question 337: What was the annual spend for the previous year on this Project?

Answer: This is a new project with no prior year spending

Question 338: Work will be onsite or remote?

Answer: Work on the project may be performed either remotely or onsite. Vendors are encouraged to propose the delivery approach that best supports effective collaboration, communication, and successful execution of the project.

Question 339: Can you please give us an extension of 1-2 weeks to submit our proposal?

Answer: Yes, the RFP has already been amended to reflect the extension.

Question 340: Could you please confirm the submission deadline for questions? The RPF contains two different dates (8/26 and 9/2). We would prefer to submit questions up until the later date, Sept 2nd.

Answer: The RFP question deadline was September 2, 2025

Question 341: Budget Constraints: Is there a defined budget range for this project? Understanding the financial parameters will help us tailor our proposal accordingly.

Answer: Funding has been allocated for this project, but a specific budget ceiling is not being disclosed at this stage. We encourage vendors to submit competitive proposals with clear and transparent pricing, including breakdowns of one-time and ongoing costs as well as any optional features.

Question 342: Stakeholder Engagement: Who are the primary stakeholders involved in this project, and how will they be engaged throughout the consulting process?

Answer: We prefer to award this contract to a single with sub-contractors (if needed). Other stakeholders are involved including Massachusetts state agencies, the MGHPCC and AICR consortium institutions, and the broader Massachusetts innovation ecosystem. Respondents should anticipate engaging across this multi-sector community to ensure that the platform reflects diverse needs and encourages wide adoption. The MA AI Hub will facilitate introductions to these stakeholders and coordinate engagement as appropriate.

Synthetic Data & Privacy

Question 343: Should the DCC support integration with federal funding sources (NIH, NSF grants) for research usage tracking and cost allocation?

Answer: This functionality is out of scope for this project phase

Question 344: Please provide a summary of any existing infrastructure we should plan on utilizing or integrating with, if applicable.

Answer: There is no existing infrastructure that vendors are required to utilize or integrate with at this stage. However, the system is expected to meet current standards and be designed for interoperability with widely used platforms and tools (e.g., Snowflake and other standard data and analytics environments), and vendors should propose approaches that ensure flexibility for future integration.

Question 345: Which domains are priority for synthetic data (healthcare, transportation, finance, robotics)?

Answer: Top priority domains include lifesciences, robotics, education, and healthcare

Question 346: Should synthetic generation support configurable privacy-utility tradeoffs (e.g., differentially private GANs)?

Answer: Yes. Synthetic data generation should support configurable privacy–utility tradeoffs.

Question 347: Is there a requirement for audit trails linking synthetic datasets back to original real datasets for validation?

Answer: Yes, traceability of synthetic data sets should be considered in scope

Question 348: Should the platform provide benchmarking tools to compare real vs synthetic datasets (distribution overlap, fidelity scores, utility metrics)?

Answer: Yes

Question 349: Should synthetic data pipelines be designed for production-grade usage or primarily for R&D and sandbox environments?

Answer: The synthetic data pipelines are not intended to operate as a full production-grade system. They should be designed primarily to support research, experimentation, and sandbox environments, while allowing for future enhancement if production-level capabilities are later required.

Question 350: Should the platform include community contribution features (dataset submission, annotation, discussion forums) from Day 1 or later?

Answer: The timing for community contribution features such as dataset submission, annotation, and discussion forums has not been finalized. Vendors should recommend whether to include these capabilities at launch or introduce them in later phases, based on best practices for usability, governance, and phased implementation.

Question 351: For synthetic data generation, is the requirement limited to tabular and text data, or does it extend to multimodal datasets (image/video/audio)?

Answer: Synthetic data generation is expected to be multimodal.

Question 352: Are there existing branding guidelines (colors, typography, logo usage) we should follow?

Answer: Branding is not required for proposal submission.

Question 353: Synthetic Data Validation: For the synthetic data capability, what are the expected criteria for validation? Is the focus on statistical similarity, utility in AI model training, privacy guarantees (e.g., k-anonymity), or a combination of these?

Answer: Validation criteria are not yet defined; vendors should recommend approaches balancing accuracy, utility, and privacy.

Question 354: Evaluation Weighting: Could you provide the relative weighting of the evaluation criteria listed in Section 4.2 (e.g., Technical Expertise 30%, Proposed Approach 25%, Budget 20%, etc.)?

Answer: The criteria identified are provided as a comprehensive set of evaluation factors and are not intended to represent a prioritized or weighted list. The MassTech will apply these criteria in a balanced and holistic manner to assess the qualifications, capabilities, and overall value of each submitter in determining the best interest of the Data Commons initiative.

Question 355: Will synthetic data generation be limited to demonstration/test datasets, or is MassTech open to full-scale synthetic derivative datasets for production AI applications?

Answer: MassTech is open to full-scale synthetic derivative datasets for production AI applications. Synthetic data generation is not limited to demonstration or test datasets, and vendors are encouraged to propose solutions that can scale to production-grade AI use cases.

User Access & Experience

Question 356: Model Training & Inference - Given that the platform will 'centralize access to data for AI development,' will users also be able to run or improve AI models on the platform—like training or fine-tuning models (for example, using GPUs for LLM inference, or running PyTorch/TensorFlow workloads)?

Answer: Yes, AI model iteration, tuning, and training are in scope for the project

Question 357: Storage Architecture - Since the platform will 'aggregate and host curated datasets' and support versioning and provenance, do you anticipate needing a scalable, tiered storage solution—like object storage for raw data and high-performance file systems for active workloads (for example, using distributed file systems or cloud-native storage services)?

Answer: Yes, a scalable, tiered storage approach is anticipated. The platform will need to accommodate multiple classes of data and workloads.

Question 358: Is there an existing design system or branding guideline that must be followed?

Answer: No.

Question 359: Is there a desire to use a Content Management System now or in the future? If so, has there been any discussion about which Content Management System to use?

Answer: There are no existing requirements for a content management system. MassTech is open to considering recommended solutions as part of the engagement.

Question 360: The RFP and Budget Template require applicants to separate direct labor costs and overhead percentages. Many highly qualified firms with the type of industry-leading AI expertise requested in the RFP provide commercial services and may not maintain the systems designed for cost-based pricing that separate direct/indirect costs. Will MassTech consider allowing firm fixed-price proposals? Or

a budget template populated with fully burdened rates, in lieu of requiring direct/indirect cost separation?

Answer: Massachusetts Technology Collaborative anticipates structuring the award in a way that aligns with best practices for both value and accountability. While a rate-based contract is an available vehicle, we are also open to fixed-price or hybrid approaches where appropriate. Our priority is to minimize the risk of budget overruns while ensuring timely high-quality delivery. We encourage respondents to propose the pricing model (rate-based, fixed price, or blended) that best matches their approach, with clear assumptions, deliverables, and mechanisms for cost control.

Question 361: Can MassTech confirm that vendors will be able to determine which tools to utilize for things like metadata management, synthetic data generation, and AI frameworks, and/or propose alternatives to the ones listed in the RFP?

Answer: The tools and examples referenced in the RFP are provided for illustrative purposes only and are not intended to be prescriptive. Vendors are expected to bring their expertise to recommend and justify the tools, frameworks, and approaches they believe are best suited for metadata management, synthetic data generation, AI frameworks, and related capabilities as part of their proposed solution.

Question 362: Will vendors need real-time data ingestion capabilities or will batch processing be sufficient?

Answer: Batch processing will be sufficient for this phase of work

Question 363: Can MassTech please detail what specific authentication protocols vendors need to support beyond SSO?

Answer: At this stage, no specific authentication protocols beyond SSO have been mandated. Vendors should propose secure, standards-based approaches (e.g., SAML, OAuth2, OIDC, or comparable methods) that align with best practices and can support the platform's governance and security requirements.

Question 364: Can MassTech please detail all third party systems vendors must be able to integrate with, including but not limited to existing billing systems and identity management systems?

Answer: There are no existing third-party systems (e.g., billing or identity management) currently in place that vendors are required to integrate with. However, the Data Commons platform should be designed with the flexibility to integrate with external systems in the future, and vendors are expected to recommend approaches and frameworks that support such extensibility.

Question 365: What existing Massachusetts government identity management systems (e.g., MassTech SSO, Commonwealth Connect) must the DCC integrate with for authentication and authorization?

Answer: There are no predefined Massachusetts government identity management systems identified at this stage for required integration. The specific authentication and authorization systems (e.g., MassTech SSO, Commonwealth Connect, or others) will be determined as part of the engagement, and vendors should recommend flexible, standards-based approaches to support future integration.

Question 366: What are the disaster recovery and business continuity requirements, including RTO (Recovery Time Objective) and RPO (Recovery Point Objective) specifications?

Answer: Specific disaster recovery and business continuity requirements, including RTO and RPO specifications, have not yet been defined. Vendors are expected to recommend approaches and propose appropriate targets based on best practices, scalability, and the criticality of the Data Commons platform.

Question 367: What level of explainability is required for AI fairness decisions, and should this integrate with existing Massachusetts algorithmic accountability frameworks?

Answer: The platform must provide interpretable, auditable fairness outputs from the start, and should be designed for future integration with Massachusetts' algorithmic accountability frameworks as they mature.

Question 368: What incident response and breach notification requirements exist, including timelines for notification to MassTech and participating agencies?

Answer: The awarded vendor will be required to maintain a formal incident response plan. Any confirmed or suspected data breach must be reported to the MA AI Hub within 24 hours, with follow-up updates as investigation and remediation progress. Participating agencies must also be notified promptly in line with applicable regulations (e.g., HIPAA, FERPA, state law). Incident response should include escalation procedures, audit logs for forensics, and post-incident reporting with corrective actions.

Question 369: What API versioning and backward compatibility requirements exist for external integrations with existing Massachusetts systems?

Answer: Backwards compatibility requirements for different types of data have not yet been defined. These requirements will be determined as part of the project, and vendors are encouraged to recommend approaches and considerations based on best practices and compliance standards.

Question 370: Do you have an existing translation solution in place or should translation solution identification and selection be included in the scope?

Answer: A translation solution is not currently in place. Recommendations for a translation solution are welcome is needed to meet accessibility standards.

Question 371: Do you have an existing CDN in place (i.e. Cloudflare)?

Answer: No. A CDN is not currently in place for the Data Commons

Question 372: Do you have any existing Identity Providers / Single Sign On (IdP / SSO) in place (i.e. Okta, Ping Identity) or should IdP/SSO solution identification and selection be included in the scope?

Answer: At this stage, no specific authentication protocols beyond SSO have been mandated. Vendors should propose secure standards and approaches that align with best practices and can support the platform's governance and security requirements.

Question 373: Is there an existing Learning Management platform in place (i.e. Blackboard, Moodle, Thinkific)?

Answer: No, there is no existing learning management platform in place

Question 374: If an existing Identity/SSO solution is available, how many Role Based Access Control (RBAC) roles apply at launch?

Answer: An SSO solution is not currrently in place and roles required at launch have not been defined. Vendors should include recommendations within scope.

Question 375: How should we address integration with existing Massachusetts systems (e.g., MassGIS, MassDOT, Public Health data sources)? Should we assume APIs will be made available?

Answer: Yes, Vendors can assume API access is available for these systems and should propose flexible integration strategies based on these API capabilities.

Question 376: Should the solution leverage existing open-source frameworks (CKAN, Dataverse, DataHub) or can vendors propose custom builds?

Answer: Vendors are free to propose either leveraging existing open-source frameworks (e.g., CKAN, Dataverse, DataHub) or developing custom builds. The decision should be guided by best fit, with proposals evaluated on scalability, interoperability, sustainability, cost-effectiveness, and alignment with the goals of the Data Commons.

Question 377: Should the platform integrate with persistent identifiers (DOI/DataCite)?

Answer: The requirement to integrate with persistent identifiers (e.g., DOI/DataCite) has not yet been defined. Vendors should recommend whether and how this should be supported.

Question 378: Should lineage tracking integrate with tools like Apache Atlas or OpenLineage?

Answer: The requirement to integrate lineage tracking with specific tools such as Apache Atlas or OpenLineage has not been defined. Vendors should recommend whether and how to leverage these or similar frameworks, based on best practices for interoperability, scalability, and governance.

Question 379: Should fairness tooling integrate with MLOps pipelines (MLFlow, Kubeflow)?

Answer: Yes

Question 380: Are there any existing websites/applications you admire for inspiration from the design standpoint?

Answer: Many examples exist. One favorable example is the Australian Research Data Commons: https://ardc.edu.au/services/research-data-australia/

Question 381: Are you open to offshore development and service delivery with onshore account management?

Answer: The MA AI Hub requires that key functions involving governance, security, compliance, and stakeholder engagement be performed by U.S.-based personnel. Vendors may propose hybrid delivery models that include offshore team members for defined roles such as software development or testing, provided this does not compromise data residency, security, or responsiveness. Proposals should clearly explain the division of onshore vs. offshore roles and how communications and compliance will be maintained. Fully onshore delivery models will also be considered and may be viewed as lower risk.

Question 382: Existing Foundation: Is this initiative starting from a completely blank slate, or are there existing systems, architectures, or datasets (at MassTech or partner agencies) that the DCC is expected to integrate with or build upon?

Answer: This initiative is starting as a new project. The platform should be designed to integrate via APIs with external systems such as other state data portals, research computing environments, and platforms like Snowflake and other commonly used solutions. Integration with MGHPCC resources is also expected as part of the broader ecosystem.

Question 383: Integration & APIs: Beyond enabling API-based access from the DCC, are there requirements for the platform to integrate via API with other external systems (e.g., other state data portals, research computing environments, etc.)?

Answer: Yes. In addition to enabling API-based access from the DCC itself, the platform should be designed to integrate via APIs with external systems such as other state data portals, research computing environments, and platforms like Snowflake and other commonly used solutions. Integration with MGHPCC resources is also expected as part of the broader ecosystem.

Question 384: Is there an offline copy of the DCC Requirements document you can share? We were not able to access the document via the link in the RFP.

Answer: The DCC Requirements document can be found at the following address: https://masstech.org/sites/default/files/2025-08/DCC%20Requirements.pdf

Question 385: Integration with Existing Systems: Are there existing systems or platforms that the consulting services will need to integrate with? If so, can you provide details on these systems?

Answer: It is anticipated that the system will integrate with other state AI platforms and capabilities including the AI Compute Resource (AICR) being implemented by MGHPCC and interoperability with the state governments use of the Snowflake architecture.