High Performance Computing Data Management Environment

project charter

Version *<1.0>*

*<06/09/2017>*

VERSION HISTORY

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Version #** | **Implemented**  **By** | **Revision**  **Date** | **Approved**  **By** | **Approval**  **Date** | **Reason** |
| 1.0 | *Zhengwu Lu* | *<06/09/2017>* | *<name>* | *<mm/dd/yyyy>* | *Initial* |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

TABLE OF CONTENTS

[1 Introduction 4](#_Toc484423888)

[1.1 Purpose of Project Charter 4](#_Toc484423889)

[2 project And Product Overview 4](#_Toc484423890)

[3 Justification 5](#_Toc484423891)

[3.1 Business Need (Value Proposition) 5](#_Toc484423892)

[3.2 Business Impact 5](#_Toc484423893)

[3.3 Strategic Alignment 6](#_Toc484423894)

[4 Scope 7](#_Toc484423895)

[4.1 Objectives 7](#_Toc484423896)

[4.2 Workflow diagrams 8](#_Toc484423897)

[4.2.1 How data flows from birth to proposed object archive? 8](#_Toc484423898)

[4.2.2 HPC Data Management Workflow with CleverSafe 9](#_Toc484423899)

[4.3 High-Level Requirements 9](#_Toc484423900)

[4.4 Major Deliverables 9](#_Toc484423901)

[4.5 Goals & Approaches 10](#_Toc484423902)

[4.6 Vision 10](#_Toc484423903)

[5 Executive Summary of Project Charter 11](#_Toc484423904)

[5.1 Project Organization 11](#_Toc484423905)

[5.2 Assumptions/Constraints 12](#_Toc484423906)

[5.2.1 Assumptions 12](#_Toc484423907)

[5.2.2 Constraints 12](#_Toc484423908)

[6 project Charter approval 13](#_Toc484423909)

# Introduction

## Purpose of Project Charter

The HPC Data Management Environment (DME) project charter formally authorizes a project, describes the business need for the project and the product to be created by the project. It provides the project manager with the authority to apply up to a certain level of organizational resources to project activities. It is created during the Initiating Phase of the project.

The intended audience of the HPC DME project charter is the Business Sponsor and Critical Partners.

# project And Product Overview

One of the most significant challenges to overcome for an effective high performance computing (HPC) support effort is effective data management, i.e., effective tracking, annotation and staging of digital datasets, accompanied with a data life cycle plan/policy for these datasets. While frequently not considered an HPC challenge or opportunity, an effective solution is needed to contain costs for stored data while increasing the scientific usefulness of data that has been created in the era of ‘big data’ where analysis of datasets can take days and total cost to store and maintain large datasets continue to tax personnel and financial resources. Without a reliable managed dataset solution, large datasets are frequently maintained in multiple copies across the physical storage in an isolated fashion, leading to an unnecessary expense as additional storage is required for analysis and storage of new data. A managed, secured, and high-availability solution will minimize the need for maintaining unnecessarily redundant copies of large datasets. Even with projected declines in the cost of physical storage, the investment in managing stored data without associated annotation will provide only minimal (at best) long-term scientific usefulness or support to advance the mission of the NCI.

Annotation and registration of large datasets is inherent for managed datasets to effectively deliver broader scientific impact and advance the mission of the NCI. Consistent with efforts already underway at the NIH within the Big Data To Knowledge (BD2K) program, annotation and registration of datasets will enable managed datasets to be of use to the community of extended and future cancer investigators. The creation and delivery of metadata and tracking utilization of datasets will provide the key insight into scientific impact for each maintained dataset.

Without an effective data management solution, the HPC effort will struggle with difficulties in staging data for analysis, recovering generated datasets, and inefficiencies created by insufficient physical storage and recomputing results that have once been completed. Therefore, we believe that:

* NCI is in critical need of advancing its core scientific and technological means of data management and services from large, diverse, distributed and heterogeneous datasets
* Large datasets are currently maintained in multiple copies across physical storage in an isolated fashion, leading to an unnecessary expense
* Annotation and registration of datasets is inherent for managed datasets to effectively deliver broader scientific impact and enable the full power of personalized medicine
* Strategically, the absence of an effective data management solution presents a barrier to supporting emerging efforts to leverage the breadth of generated datasets for use in development of computationally and data intensive predictive models as well as efforts to utilize cloud resources for collaboration and analysis.

# Justification

## Business Need (Value Proposition)

Frederick Core Sequencing Facility (SF) is our critical business partner and has been engaging into every step of our efforts:

* SF needs to register CSAS project, run, sample and associated metadata
* SF needs to register, store, transfer and share scientific data/dataset or files with investigators or collaborators
* SF needs to leverage their metadata from existing LIMS
* SF needs to discover dataset/data object based on its metadata to securely share with collaborators
* SF have adopted Globus/GridFTP technology to transfer and share large sequencing datasets with CCR collaborators
* SF needs to have a permanent archive of their annotated datasets vs pieces of many data islands to share with collaborators and enable metadata-based data life cycle

With SF adoption of the HPC DME solution, our expectation is that the APIs, code-base, documentation, training material and toolset will be made available for all NCI, FNL collaborators, investigators in assisting scientific data management activities across various laboratories or DOC groups/team involved in generating, analyzing and utilizing scientific datasets. Similar or extended use cases will be verified and/or confirmed through strong collaborations and engagement with the end user audiences.

## Business Impact

* Assess scientific data management, storage islands, enhance share and analytic improvement opportunities at NCI
* Identify current data governance processes and ownership within NCI and its critical partners.
* Champion the creation of an enterprise dataset archive, its affiliated metadata catalog and APIs as a foundation for cancer research enterprise competitiveness. This object datasets archive is a critical step in creating a robust analytic infrastructure. The true value of the object archive and its metadata repository is to organize data, provide links across disparate data sources (so the analysts don’t have to), and provide access so analysts and [scientists](https://www.healthcatalyst.com/physician-reporting-secret-useable-engaging-reports) can “fish for themselves.”
* Improve patient care and deliver better outcomes with a new data services model that supports effective scientific data management to assemble and coordinate data from across the organization. This streamlined model greatly reduces lead time, enabling analysts to provide strategic insight from the disparate data collected across various systems.
* Minimize the management of redundant datasets
* Provide necessary information for selecting the appropriate datasets of interest for subsequent analysis
* Provide the foundations for effectively staging of datasets used as input to HPC-enabled analysis, modeling and simulation.
* Provide the foundations for effective recovery of datasets created in the analysis, modeling and simulation efforts
* Provide the ability to minimize costly recalculation by effectively tracking valuable results once generated.
* Strategically position NCI to utilize datasets in the development of predictive computational models
* Strategically position NCI intramural investigators to use cloud resources for future analysis

## Strategic Alignment

| **Goal** | **Project Response Rank** | **Comments** |
| --- | --- | --- |
| *Scale*: **H** – High, **M**- Medium, **L** – Low, **N/A** – Not Applicable | | |
| **Organization Strategic Goals:** | | |
| *We need sophisticated computational models to understand patient response, methods of*  *resistance, and to integrate pre-clinical model*  *data* | H |  |
|  |  |  |
| **OPDIV/STAFFDIV Strategic Goals:** | | |
| *[Enter OPDIV/STAFFDIV goals]* |  |  |
|  |  |  |
| **Department of Health and Human Services (DHHS) Strategic Goals:** | | |
| *[Enter Health and Human Services Goals]* |  |  |
|  |  |  |
| **DHHS IT Goals:** | | |
| *[Enter Health and Human Services IRM Goals]* |  |  |
|  |  |  |
| **President’s Management Agenda (PMA) Strategic Goals:** | | |
| *Precision Medicine Initiative* | L | Precision medicine will lead to fundamental understanding of the complex interplay between genetics, epigenetics, nutrition, environment and clinical presentation and direct effective, evidence-based prevention and treatment. |
|  |  |  |

# Scope

## Objectives

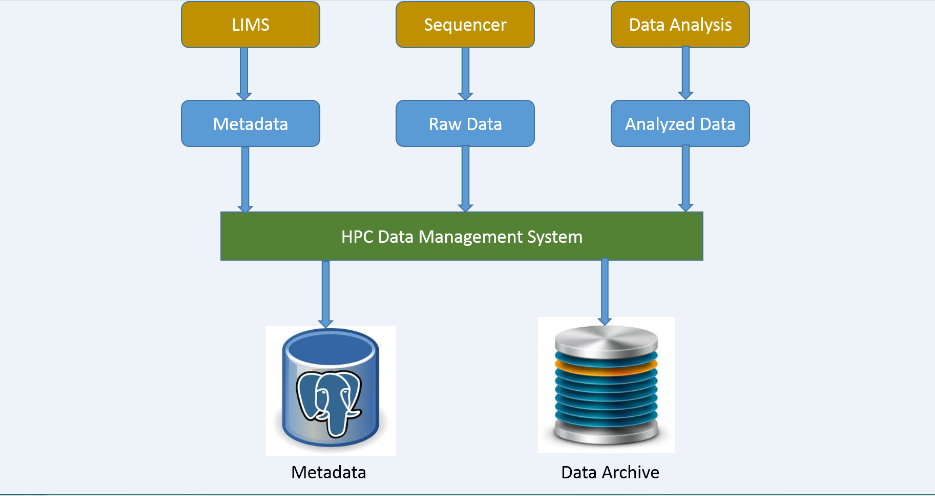
CBIIT, with input from the Office in Scientific Operations (OSO) in Frederick, has identified a requirement to alter the operations and scope of support that are currently offered to NCI for IT services and Informatics support from The National Cancer informatics Program. They recognize the need to: i) establish direct relationships with the Divisions and Centers throughout NCI, ii) gain the trust of those constituents, and iii) re-architect the scientific computing and enterprise IT infrastructure, as well as the IT support services and informatics support offered by CBIIT and the Frederick National Laboratory to better match the requirements of NCI.

It is critical that the prototyped APIs and web front end address the prioritized use cases to meet the user workflow needs.

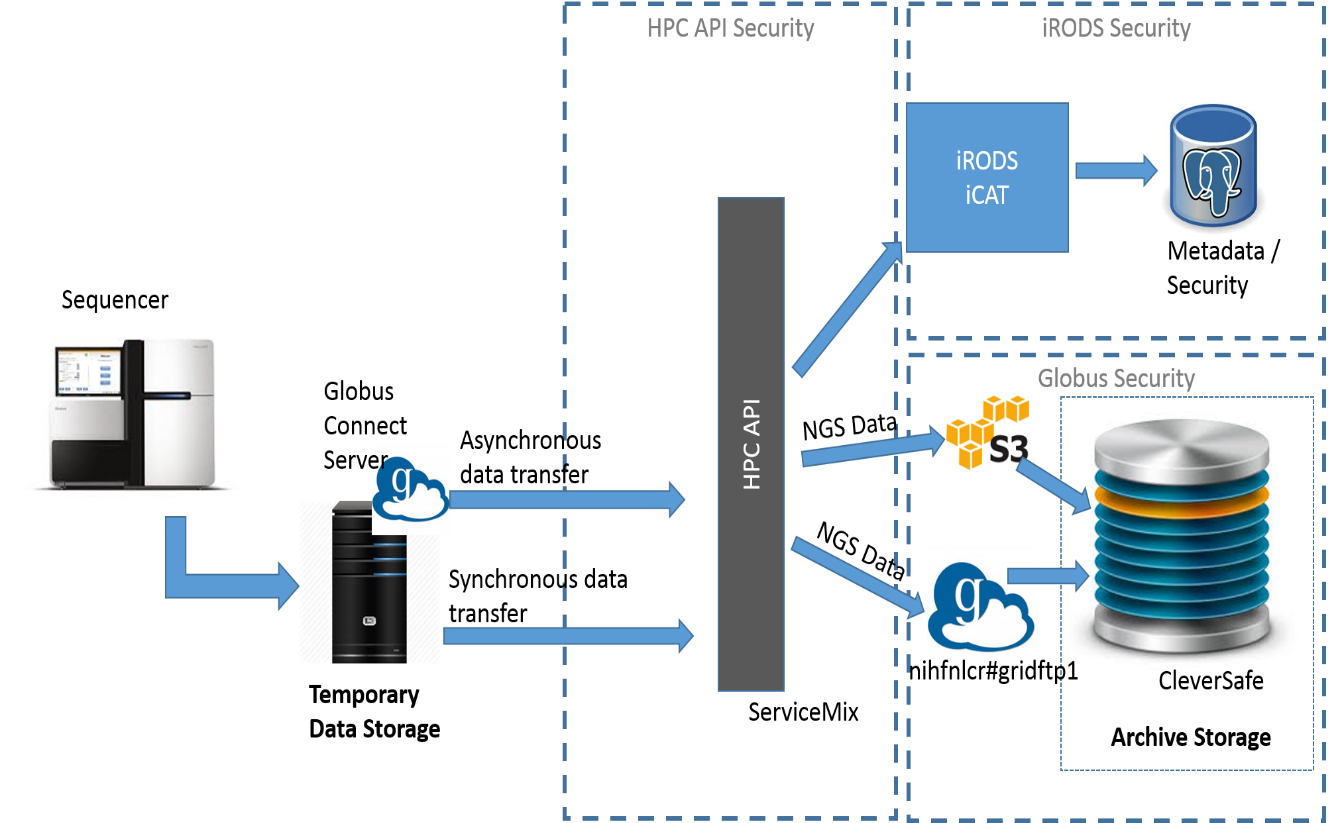
* Establish GridFTP services to support general transfer of large datasets without requiring physical mounting.
* Establish a pilot dataset registration system to associate a label with a given managed dataset. The System will capture extensible metadata including security and access requirements for the managed dataset. Metadata includes but not limited to: How the dataset was generated, when it was generated, where it was generated, present access method and information to obtain a full copy of the original dataset. System will also be flexible to support either a) export of metadata to a future system; b) development of service APIs to support interrogation by secondary systems; or c) both (a) and (b).
* Establish easy-to-use methods for providing annotation information and creating reports of managed datasets, including utilization of a modest controlled vocabulary related to high-use criteria for searching datasets.
* Establish a high-reliability storage model for underlying datasets included into the dataset registration system
* Obtain utilization statistics for managed datasets.
* Provide required system administration support for data management services
* Provide development and general fixes to existing implementation
* Design and introduce Web-based interfaces for submitting, retrieving and locating data by metadata characteristics.
* Design and implement enhancement features per prioritizations
* Pilot and implement DevOps and required operational modifications to support policies around HPC data management services environment
* Extend implementation to support placement of data objects in cloud (based on requirements to be provided)
* Implement and support shared development and collaborating with NIH/CIT and others in updating and extending the code-base
* Produce updated assessment of comparable/compatible technologies.
* Participate in the evaluation of the CGC pilots, focusing on analyzing and testing of the APIs and web front ends ability to handle workflow needs in handling of big data.

## Workflow diagrams

### How data flows from birth to proposed object archive?



### HPC Data Management Workflow with CleverSafe



## High-Level Requirements

The following table presents the requirements that the project’s product, service or result must meet in order for the project objectives to be satisfied.

| Req. # | I Requirement Description |
| --- | --- |
| 1. | HPC DME shall support NCI LDAP authentication |
| 2. | HPC DME shall support large data file upload/download |
| 3. | HPC DME shall be integrated with archive storage |
| 4. | HPC DME shall enable support and integration with metadata repository |
| 5. | HPC DME shall support user organizations of data objects and collection [hierarchies](https://www.google.com/search?q=hierarchies&spell=1&sa=X&ved=0ahUKEwifuMOAtN3OAhVoIcAKHRjtCVYQBQgdKAI&biw=1920&bih=880) |
| 6. | HPC DME shall provide batch utility tools for batch uploading large number of files |
| 7. | HPC DME shall provide common APIs for users to call; or execute single update or search functions |
| 8. | HPC DME shall provide storage utilization and patterns. |
| 9. | HPC DME shall be able to scale up, extend and support S3 APIs or cloud based technologies. |
| 10. | HPC DME shall enable loosely coupled, highly customizable, autonomous, composable and discoverable APIs supporting agile development approach. |

## Major Deliverables

The following table presents the major deliverables that the project’s product, service or result must meet in order for the project objectives to be satisfied.

| Major Deliverable | I Deliverable Description |
| --- | --- |
| 1. | HPC DME Design/Integration |
| 2. | HPC DME Requirements Definition |
| 3. | HPC DME DR and Monitoring Plan |
| 4. | HPC DME Server API specification |
| 5. | HPC DME System Administration Guide |
| 6. | HPC DME User Guide |
| 7. | HPC DME General Training |

## Goals & Approaches

* Provide core capabilities to get started , but extensible to accommodate future need
* Implement/enhance HPC data management core APIs based on iRODS iCAT and Jargon core APIs – The unified REST APIs will be agnostic of physical storage medium or device being utilized
* Implement HPC DM command and batch utilities with no interference of currently SF adopted Globus workflow
* A highly flexible and reliable storage model for underlying collections and data objects
* Data virtualization with multiple storage types
* Configurable metadata policies for validation
* Secure APIs enforcing authentication and authorization
* Data discovery through descriptive metadata
* Data sharing through REST push and pull methods
* Integration with iRODS, Cleversafe and Globus Transfer API
* Command and batch utilities to register collections, data objects and update permissions
* Configurable security layer extending iRODS security implementation
* Access to Globus, iRODS and Cleversafe through system service account
* Pluggable data transfer implementations (S3, Globus, iRODS)
* Assess technology to focus on adding value for identified use cases
* Re-use and enhance vs invent new tools to address SF needs
* Enhance, extend and collaborate as engagement with other groups goes on.

## Vision

* By introducing the core HPC DME APIs, related batch utility toolsets, and Web GUI applications and features, data collected at NCI will help achieve strategic goals.
* By creating and maintaining an object archive repository and associated metadata catalog, value added services or predictive modeling will be tapped and utilized for potential new therapeutic interventions or other competitive advantages
* Improve and enhance community involvement and knowledge sharing via extending the HPC DME services, integrating with an existing Cloud service or exposing through a Cloud based host.

# Executive Summary of Project Charter

## Project Organization

The project organizational structure for the NCI HPC Data Management Environment effort, shown in Figure 1 below, is designed to facilitate collaborative management of the interdependent cross-agency business and technical activities needed to complete this project.

NCI CBIIT, as owner of the High Performance Computing Data Management Environment initiative for which APIs, tools, solutions and related operational support will be targeted and implemented, will collaborate and partner with the Leidos Biomedical Research Inc (LBR) HPC team and its subcontractor in collecting business driven use cases, alignment with NIH BD2K advancement, and agile development and implementation of the object archive with associated infrastructure and capital support.

LBR, specifically DSITP, will be responsible for project management, procuring and managing necessary subcontractor or consultant resource in providing needed expertise, technical review, quality verification and confirmation of the APIs, toolsets or developed features in supporting core use case scenarios by working closely with critical business partners or groups, procurement of necessary storage hardware and software, and coordinating with NCI/NIH CIT group in piloting, collaborating and adopting HPC DME APIs, toolsets, code-base, or solutions.



***Figure 1. Project Organization Chart***

## Assumptions/Constraints

### Assumptions

* Agreement on objectives: A consensus agreement between all stakeholders regarding the objectives, vision, goals and approach for this project has been or will be reached.
* LBR DSITP storage resources available and committed: NCI CBIIT infrastructure core team and other resources as necessary will be available and able to successfully support the deliverables of this project.
* Other priorities will not delay deliverables: Other IT/IS initiatives will not delay the delivery of this project deliverables.
* Financial Resources are available: Financial resources will be available for all necessary aspects of this project.

### Constraints

* Delays on the CleverSafe storage purchasing process due to the third party vendor-manufacturer involvement and subcontracting paperwork.
* Because of the limited and conflicting availability of key stakeholders in LBR ITOG storage and system resource and NCI CBIIT infrastructure, ample notice and lead time may be required or expected when their participation in meetings, reviews, tasks execution or other engagements is required
* Vendor delays on availability of required connector or plugins
* Specialized storage design and implementation expertise/support may be procured under a separate subcontract through LBR
* NCI HPC DME piloting, roll-out/implementation (DevOps), or extension to Cloud must be coordinated and communicated through all NCI HPC critical stakeholders, vendors and impacted parties including our business partners to minimize operational and integration issues from implementation perspectives.

# project Charter approval

The undersigned acknowledge they have reviewed the project charter and authorize and fund the HPC DME project. The undersigned herby give the project manager the authority to apply the approved level of organizational resources to project activities. Changes to this project charter will be coordinated with and approved by the undersigned or their designated representatives.

[List the individuals whose signatures are desired. Examples of such individuals are Business Sponsor and Project Manager. Add additional lines for signature as necessary. Although signatures are desired, they are not always required to move forward with the practices outlined within this document.]

|  |  |  |  |
| --- | --- | --- | --- |
| Signature: |  | Date: |  |
| Print Name: |  |  |  |
| Title: |  |  |  |
| Role: |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Signature: |  | Date: |  |
| Print Name: |  |  |  |
| Title: |  |  |  |
| Role: |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Signature: |  | Date: |  |
| Print Name: |  |  |  |
| Title: |  |  |  |
| Role: |  |  |  |