Data Storage, Toolspace, Access, and Analytics for biG-data Empowerment (DataSTAGE) Implementation Plan

V1.0 - 20190426

DataSTAGE

DataSTAGE Implementation Plan - DRAFT

V1 - 20190426

Document Status

Version

V1.0

Approvals

Signatures presented below denote review and approval of the DataSTAGE Implementation Plan. These approvals are given based on the understanding that the Implementation Plan, and the information herein, will be revised at regular periods over the course of the program. It is the responsibility of the Principal Investigator (PI) of each funded team and select NHLBI program staff to add their name(s) in the indicated space below.

Approved Date

4/26/2019

PI Approvals:

PI	Team	Approval Date
Robert L. Grossman	Calcium+	03/27/2019
Paul Avillach/Isaac Kohane	Carbon+	03/26/2019
Ashok Krishnamurthy	Helium+	03/26/2019
Brandi Davis-Dusenbery	Xenon+	04/01/2019

NIH Approvals:

Responsible Person	NIH NHLBI DataSTAGE Role	Approval Date
Jonathan Kaltman	Program Manager	4/26/2019
Alastair Thomson, NHLBI CIO	Information Security	04/17/2019

Next Review Date

4/26/2020

Document Owner

STAGECC

Revision History

Date (YYYYMMDD)	Version Number	Revision Reviewed/ Approved By	Brief Description of Change
20190110	V0	N/A	Draft document created
20190206	V.0.1	Stan Ahalt	Content update: all sections
20190214	V0.2	Paul Avillach (Carbon team)	Added i2b2/tranSMART platform and PIC-SURE metaAPI as the “gold master” for clinical data in DataSTAGE
20190226	V0.3	Rebecca Boyles	Changes accepted, and comments addressed
20190314	V0.3	Marcie Rathbun	At the end of section 3.2, added link to Operationalization document: NHLBI DataSTAGE 60 Day o16n Plan v1-2
20190426	V1.0	NHLBI	V1.0 reviewed and approved by NHLBI Links, graphics, & editing updates [Marcie]

TABLE OF CONTENTS

INTRODUCTION 5
PURPOSE OF THE IMPLEMENTATION PLAN 5
BACKGROUND 5

OVERVIEW 5
KEY TERMS AND CONCEPTS 6
PROGRAM ORGANIZATION 6
Consortium Groups 6
Coordination of Activities 7

DATASTAGE PLATFORM OVERVIEW 9
SYSTEM DESCRIPTION 9
SYSTEM DEVELOPMENT 11
User Narratives 2019-2021 12
Features, Epics, and User Stories 12
Cross-Team Development Coordination 12

TRAINING, USER ENGAGEMENT, AND ASSESSMENT 13
TRAINING 13
Ambassador Program 14
Beta-User Training 14
USER ENGAGEMENT 15
Community Outreach 15
ASSESSMENT 15

SUMMARY 16

REFERENCE DOCUMENTS 16

REFERENCES **Error! Bookmark not defined.**

1 INTRODUCTION

1.1 PURPOSE OF THE IMPLEMENTATION PLAN

The DataSTAGE Implementation Plan describes the process by which the DataSTAGE Consortium will incrementally progress towards the vision of the program described in the DataSTAGE Strategic Framework. The Implementation Plan will enable the teams to decompose the strategic vision into concrete steps and define measures of completion for each step. Additionally, the Implementation Plan and Strategic Framework will focus the Consortium on the steps necessary to execute on the DataSTAGE strategy.

1.2 BACKGROUND

This document outlines how the various elements from the planning phase of the DataSTAGE project will come together to form a concrete, operationalized DataSTAGE platform. The platform will offer the ability to perform novel science and access an unprecedented array of data to a diverse set of users. The platform will advance groundbreaking research and significant advances in medicine.

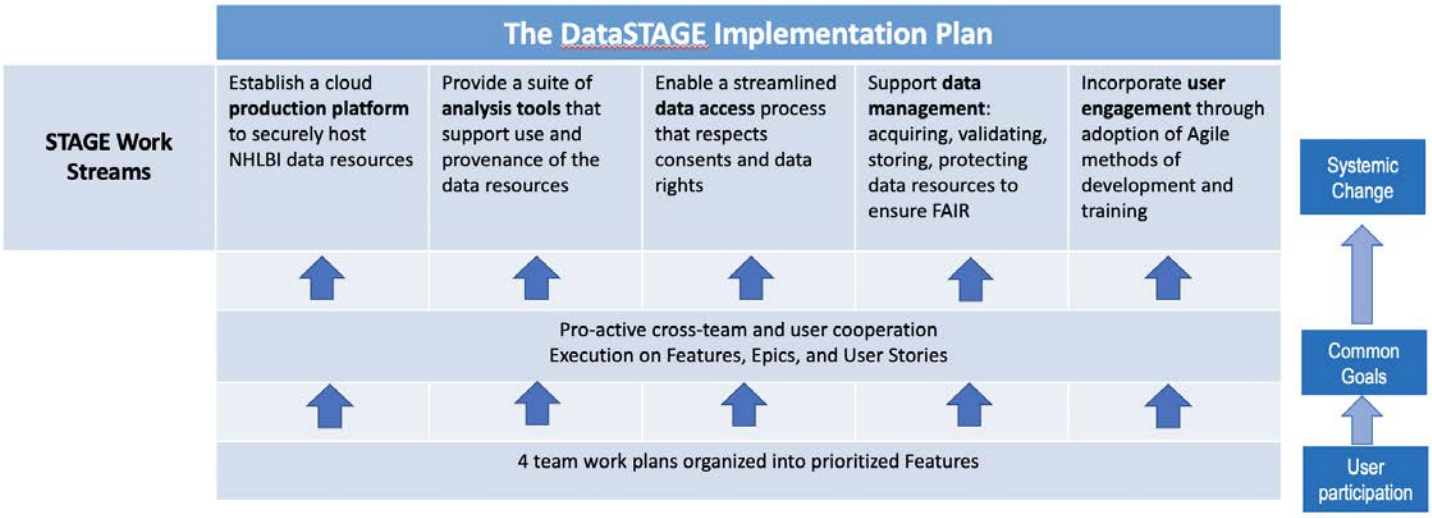
The Implementation Plan, coupled with the Project Management Plan, establishes priorities and accountabilities for resource use. The DataSTAGE Project Management Plan describes how DataSTAGE will execute, monitor, and control work towards deliverables within the program. To create project priorities and transparency, the Implementation Plan uses the following guiding principles:

- Maximizing availability of resources;
- Positive impact on the NHLBI mission;
- Responsible stewardship of funds;
- Utilization of technologies that maximize data security and integrity;
- Implementation of cost-effective solutions; and
- Consistency with the DataSTAGE Consortium Charter and values.

2 OVERVIEW

The DataSTAGE Consortium (DSC) is a collection of teams and stakeholders working to deliver on the common goals of integrated and advanced cyberinfrastructure, leading-edge data management and analysis tools, FAIR data, and HLBS researcher engagement. The DSC takes an Agile development approach towards implementation to be flexible and responsive to user needs and react to user feedback. Accordingly, work within the Consortium is described at various levels of detail with a focus on collaboration with the user community.

The organization of the program will be coordinated according to objectives that are captured in Work Streams, which are further decomposed into Features, Epics, and User Stories (see figure below).



Key functionalities in DataSTAGE are organized via technological groupings called Work Streams. The figure above illustrates how the DataSTAGE teams' activities will roll up into these Work Streams. Work Streams will be comprised of Features, which will be comprised of large-scale Epics, which can be broken down into smaller-scale User Stories. Coordination along this framework will support the DataSTAGE teams working in a coordinated manner towards common goals. User Narratives are an orthogonal construct to Work Streams, but are critical in the integration of user participation into the development cycle. Project milestones are defined and tracked via the delivery of Features, Stories, and Epics.

2.1 KEY TERMS AND CONCEPTS

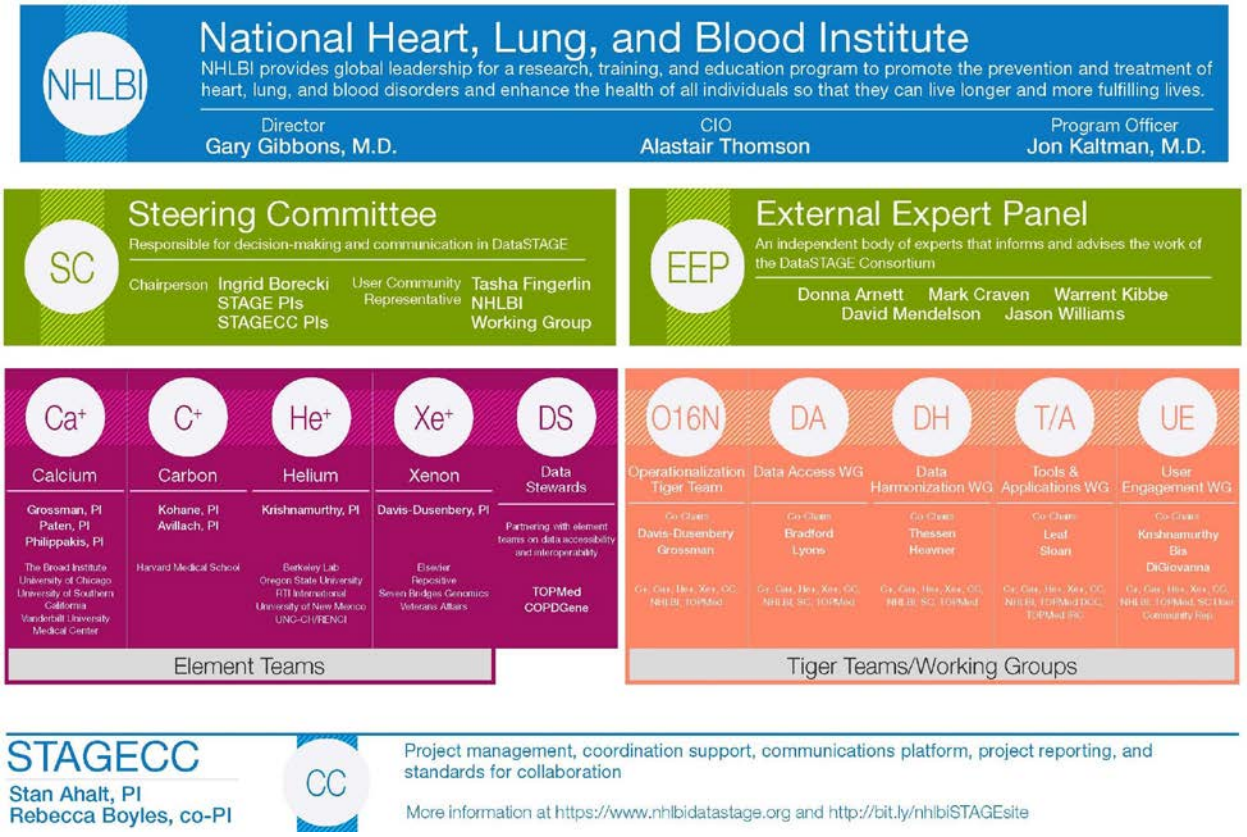
In the DataSTAGE program, the following key terms are used:

- **User Narrative**: A description of a user interaction experience within the system from the perspective of a particular persona. Example: An experienced bioinformatician wants to search TOPMed studies for a qualitative trait to be used in a GWAS study.
- **Feature**: A functionality at the system level that fulfills a meaningful stakeholder need. Example: Search TOPMed datasets using the i2b2/tranSMART platform.
- **Epic:** A (very) large User Story described at the program level that can be broken into executable stories. Example: i2b2/tranSMART is accessible on DataSTAGE.
- **User Story**: An item that describes a requirement or functionality for a user. Example: A user can access i2b2/tranSMART through an icon on DataSTAGE to initiate a search.
- **Work Stream**: A collection of related features; orthogonal to a User Narrative. Example: Work Streams impacted by the above User Narrative include production system, data analysis, data access, and data management.

2.2 PROGRAM ORGANIZATION

Consortium Groups

The DataSTAGE program is composed of several groups who each bring various resources towards executing the vision of DataSTAGE. The organizational chart below displays the teams and their responsibilities.



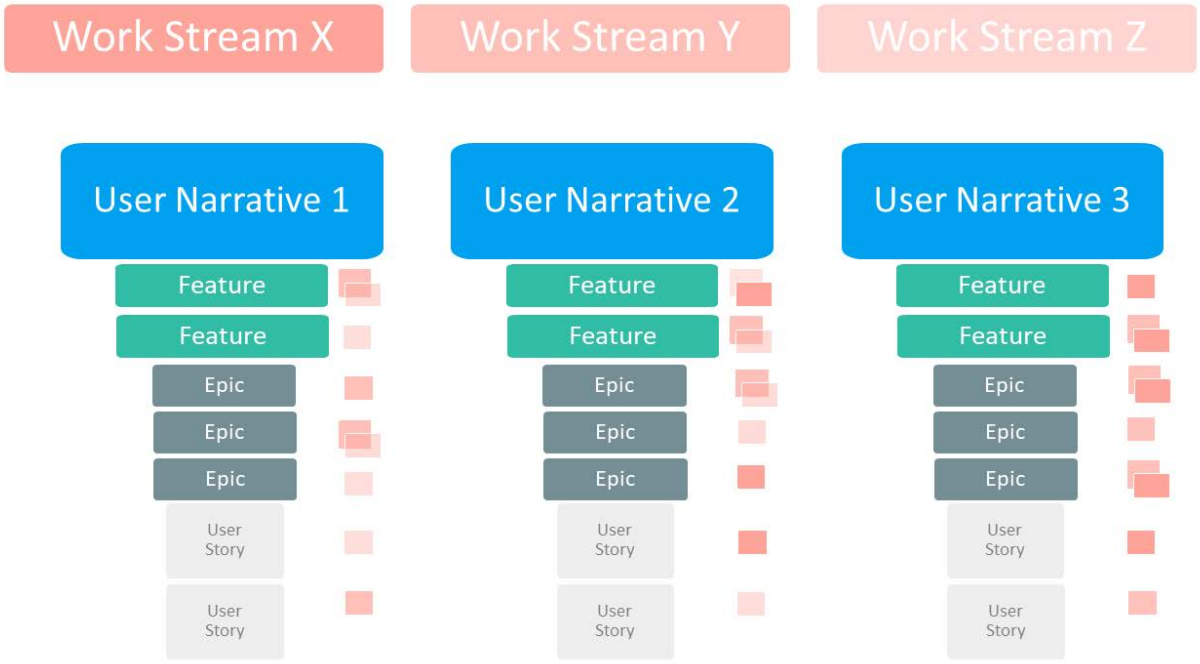
The DataSTAGE Coordinating Center (STAGECC), in collaboration with NHLBI, develops and maintains the Strategic Framework, Implementation, and Project Management Plans. All members of the Consortium are periodically invited to provide feedback on these plans to the STAGECC, with a particular focus on integrating feedback from the Data Stewards and users. The draft documents and any significant changes are reviewed by the Steering Committee as well as the External Expert Panel. The Ca+, C+, He+, and Xe+ teams are responsible for collaborating to deliver Features, Epics, and User Stories and advance the DataSTAGE platform. Additional details on the membership, roles, and responsibilities of each group can be found in the Project Management Plan.

Coordination of Activities

Initially executed by four teams, meeting the goals of DataSTAGE requires intense and ongoing collaboration to create cyberinfrastructure, tools, processes, and a community of practice. The software development teams within the DataSTAGE Consortium will be largely self-organizing around Epics, which the STAGECC will coordinate to ensure synchronization across shared Features. Multiple Features will commonly compose a User Narrative. Successful completion of work will be measured against the ability for a user to complete the work outlined in a User Narrative.

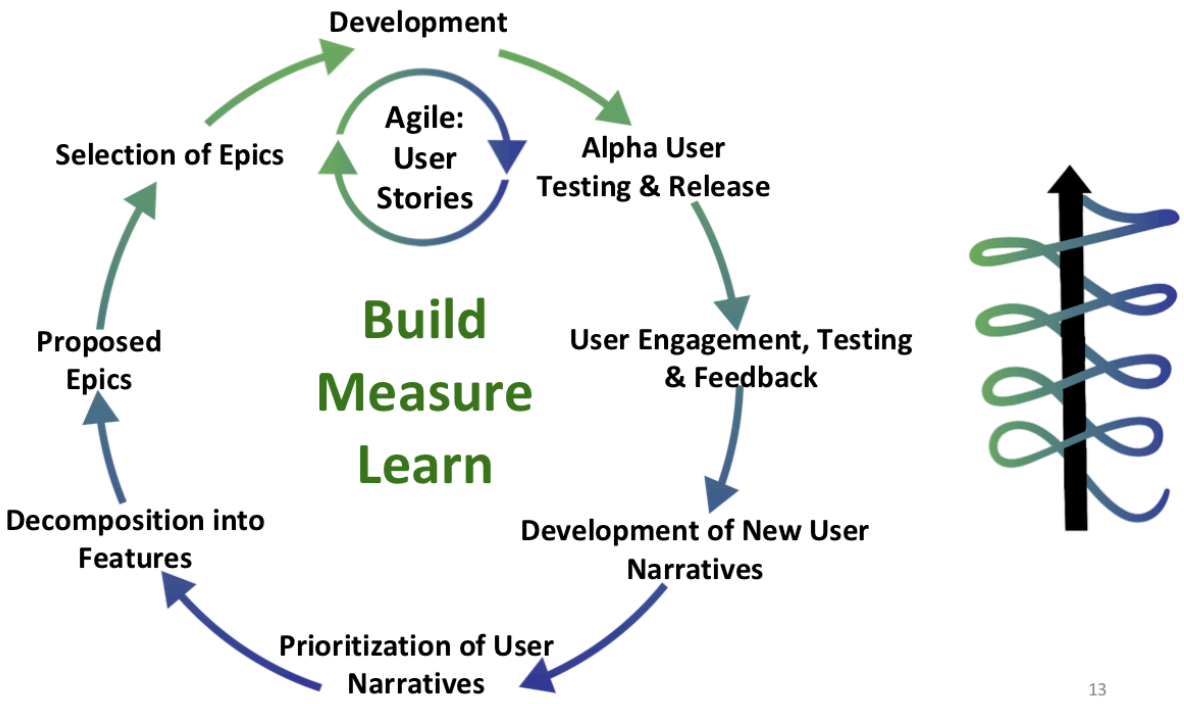
The ability for a user to complete a User Narrative on the DataSTAGE system will indicate meaningful progress towards completion.

Independently, Features, Epics, and the more granular User Stories can be mapped to Work Streams, which are useful for reporting on aggregations of specific types of work, as shown in the figure below.



DataSTAGE maintains a Consortium glossary of terms that is regularly updated in the STAGE-RFC-2 DataSTAGE Strategic Planning Nomenclature.

A cyclical evaluation and revision of the DataSTAGE User Narratives will be critical to the execution of the DataSTAGE Agile program. Regular collection of user feedback and needs will feed into the development process and be represented in new or revised User Narratives that will be prioritized in coordination with NHLBI. This continued and organized refinement of priorities for Consortium development work will support close coordination and ground the DataSTAGE program in the needs of the user community.



3 DATASTAGE PLATFORM OVERVIEW

3.1 SYSTEM DESCRIPTION

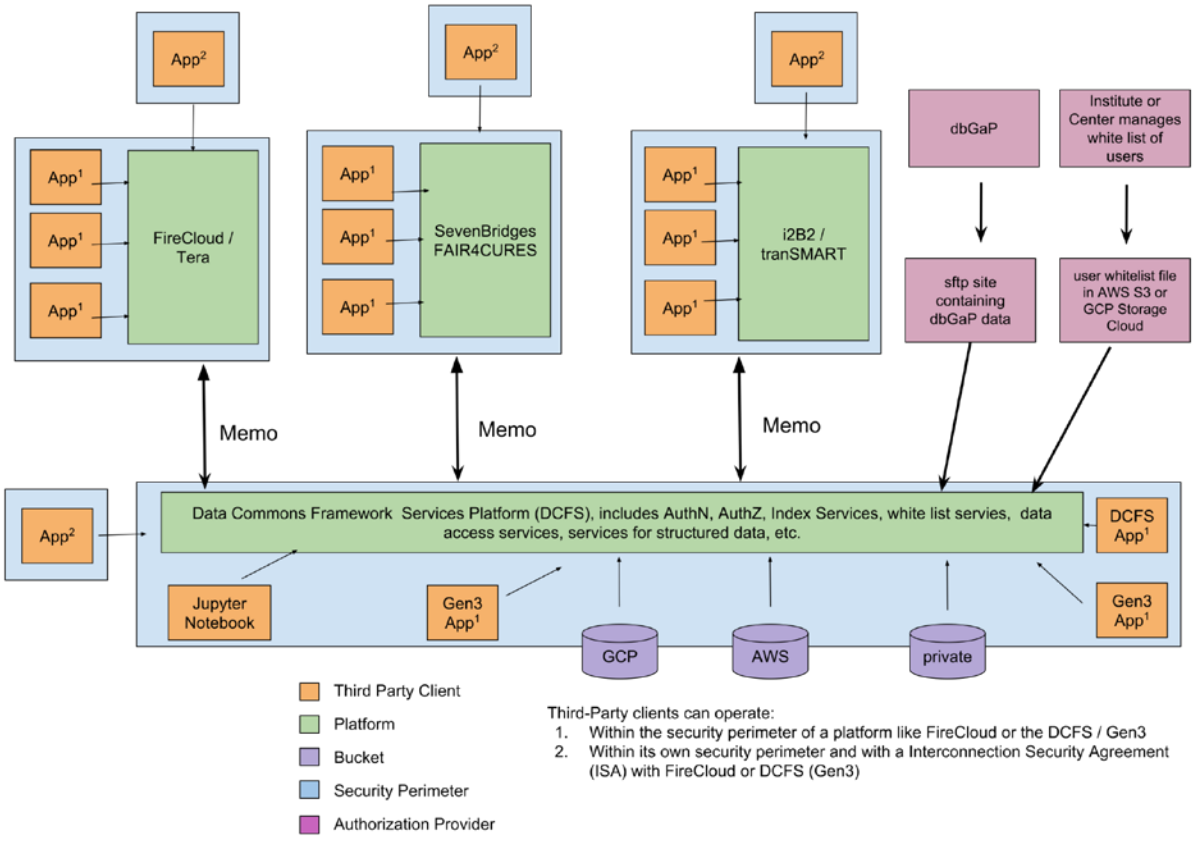
Inherent in the approach to the system design is the recognition that the current state of data and computational resources places onerous limits on the HLBS research community. Examples of limitations include an inability to execute arbitrary code, inability to access and work on very large data (e.g., TOPMed CRAMs) due to technical constraints, inability to search on one platform and execute on another, difficulties for groups of investigators to share controlled-access data and work together in a common workspace, as well as a laborious, several month process for a researcher gaining access to data.

The DataSTAGE architecture provides an early cyberinfrastructure to researchers as quickly and responsibly as possible with an eye towards addressing the above limitations. DataSTAGE will balance early delivery with ambitious goals by extending functionality through phased rollouts.

To accomplish this goal, the Consortium will abide by the below design principles:

- Meet user needs and incorporate feedback
- Leverage existing tools and infrastructure, when feasible
- Duplicate functionality when intentional and reasonable
- Architect interoperability with relevant systems
- Encounter a seamless experience, regardless of underlying components
- Leverage cost-advantageous cloud resources
- Support scalability and extension of functionality
- Have an early impact on computational-driven HLBS science
- Enable easy access to applications and tools for users across DataSTAGE
- Provide systems security for hosting identifiable data

Applying these design principles, our initial architecture of the DataSTAGE platform is pictured in the figure below. The teams will leverage the Data Commons Framework Services (DCFS) of Gen3 to provide critical infrastructure, common security, data access services, and the genomic data gold master. The DCFS is a set of software services designed specifically to support this kind of Data Commons platform. The DCFS is powered by the Gen3 platform and were initially developed to support the National Cancer Institute's (NCI) Genomic Data Commons (Grossman, 2018). The i2b2/tranSMART platform will be the clinical data gold master database leveraging the PIC-SURE metaAPI. These data services will make use of the NIH STRIDES partnerships that offer NIH investigators cloud services and storage at discount pricing to support research (NIH, 2018).

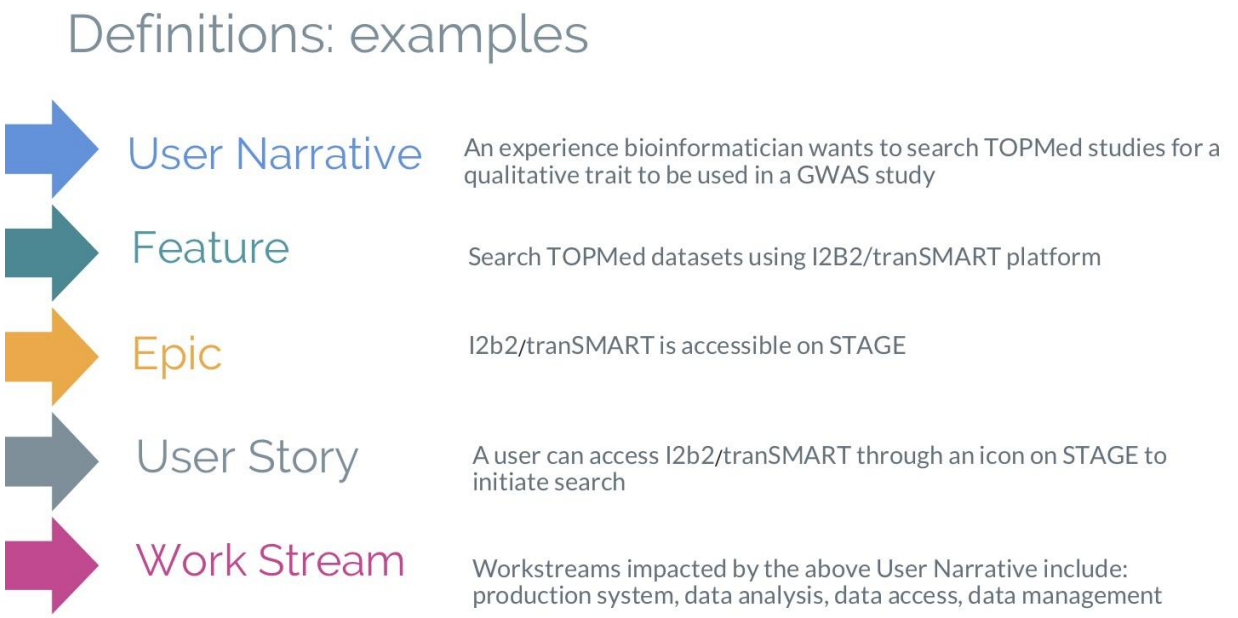


The DCFS will include authentication and authorization services and digital object globally unique IDs for indexing. The current FireCloud, Seven Bridges, and i2b2/tranSMART platforms will establish appropriate memos for interoperation with the DCFS. These memos are means through which groups will formalize cooperation with one another to develop interoperability solutions that meet functional, technical, and security/compliance requirements.

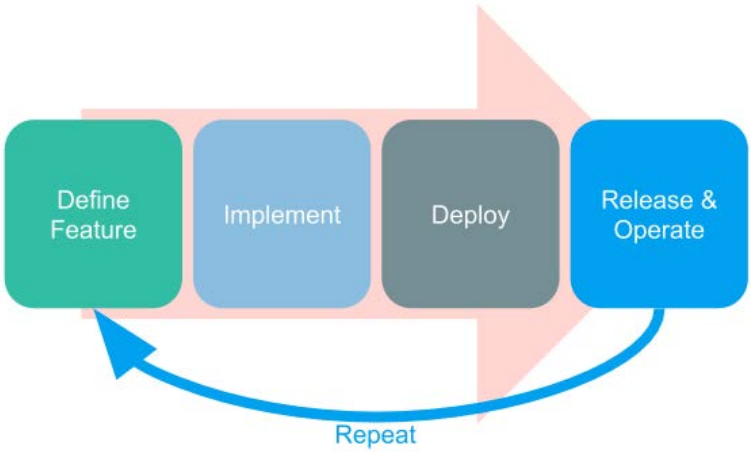
DataSTAGE will be extended through the integration of third-party applications. There are a number of possible models in which a third-party application can operate within the DataSTAGE platform. The terms of operation for these applications are being developed collaboratively between the Tools and Applications Working Group and the Operationalization Tiger Team.

Initial architecture approaches for establishing the DataSTAGE system are further described in the NHLBI DataSTAGE 60 Day o16n Plan.

3.2 SYSTEM DEVELOPMENT



User Narratives 2019-2021



The phased development of the DataSTAGE platform will be orchestrated through the collection, prioritization, and execution of User Narratives. User testing will be performed against these narratives to ensure appropriate completion. User Narratives offer an opportunity to engage potential users in the development process. As these users begin to work within DataSTAGE, they will identify additional User Narratives that are needed to advance their research. These User Narrative updates will be reflected in regular revisions to the Strategic Framework and Implementation Plan to be reflected in future development efforts.

Features, Epics, and User Stories

The relationship between Features, Epics, and User Stories is hierarchical. A Feature is a service that fulfills a stakeholder need. Releases are managed at the level of the Feature and will be coordinated by the STAGECC in conjunction with the development teams and stakeholders. To support prioritization and acceptance testing, Feature descriptions include their benefits and criteria for acceptance.

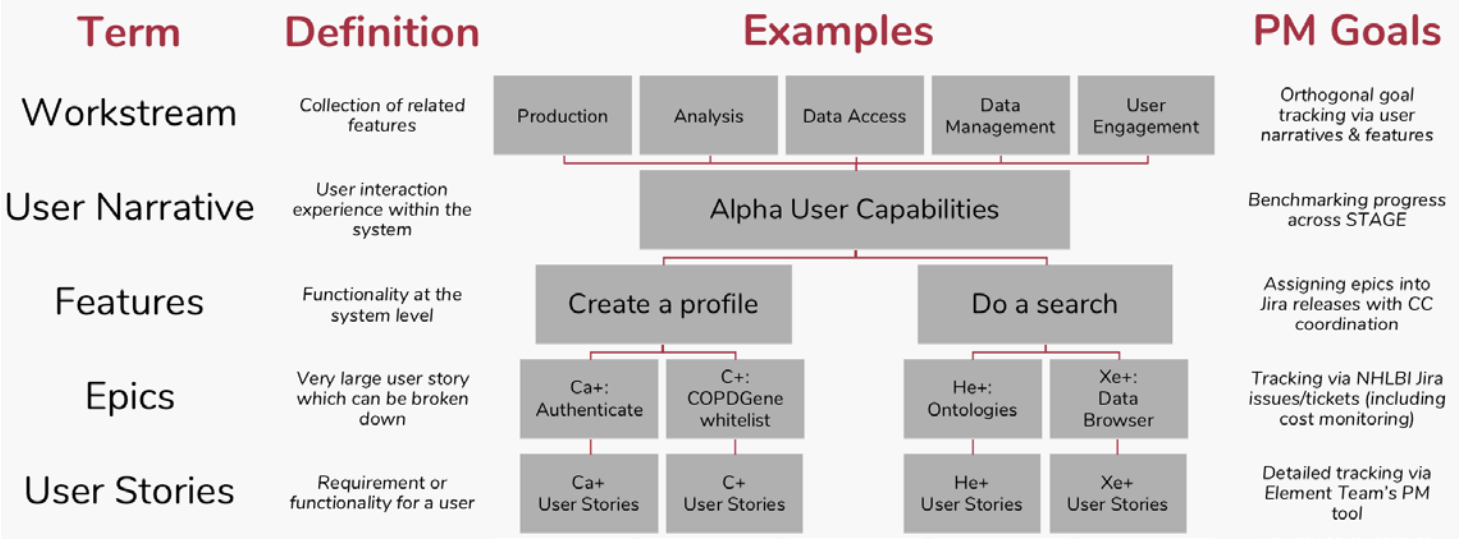
Initial Groups
Operationalization Tiger Team
Tools and Apps Working Group
Data Harmonization Work Group
Data Access / UX-UI Working Group

Features conceptually map to Work Streams, which are groupings of technologically-related work. In DataSTAGE, these Work Streams map to DataSTAGE Working Groups and Tiger Teams who will help to gather additional information and inform the creation of Features. These small Working Groups and Tiger Teams will center on collaborative opportunities within the Work Stream areas. The STAGECC will work with the teams and appropriate Working Groups and/or Tiger Teams to coordinate activities, deliverables, and releases. The initial Working Groups and Tiger Teams are listed and reflect the current focus of DataSTAGE efforts. By design, these groups will evolve over the course of the project.

Cross-Team Development Coordination

Coordinating across the Consortium towards creating a cutting-edge cyberinfrastructure requires a framework to support individual development as well as collaboration. The DataSTAGE program will work within the framework below to communicate with teams and track outcomes and dependencies.

A User Narrative is a description of how a particular user will interact with the system, often crossing Work Streams and including many different types of Features.



As briefly described in the previous section, the STAGECC will coordinate across teams, and for the purpose of software releases, at the Feature level of development. Individual teams will manage development work according to Epics; Epics are best described as very large User Stories. High-level requirements are gathered at this level and are further refined by teams into User Stories. DataSTAGE teams will report the cost to NHLBI at the Epic level. The individual teams will track resources and schedules at the User Story level. Additional details on reporting can be found in the Project Management Plan.

Important elements of building a successful platform are user experience and feedback. STAGECC intends to develop the training, user engagement, and assessment strategy to ensure efficient onboarding of users, training, and feedback leading to streaming software and solution improvement. Refer to section “Training, User Engagement, and Assessment” below for more details.

Phase 1 of the System Development is further detailed in the NHLBI DataSTAGE 60 Day o16n Plan v1-2.

4 TRAINING, USER ENGAGEMENT, AND ASSESSMENT

The training, engagement, and assessment strategy within DataSTAGE will be a phased approach intentionally developed towards the needs of particular user communities as defined by the User Narratives. By approaching training in phases, we anticipate the establishment of early DataSTAGE Ambassadors who are invested in the development of the DataSTAGE platform and are active contributors of feedback to the development teams.

4.1 TRAINING

As DataSTAGE evolves as a program and a system, our approach to training will also evolve. The assessment of programs will lead to modifications as we take an Agile approach to training.

Ambassador Program

DataSTAGE will identify a small number of expert users and/or those users with unmet, urgent needs, initially from the TOPMed and COPDGene programs, to act as DataSTAGE Ambassadors. These Ambassadors will likely represent the personas featured within the priority User Narratives. We envisage DataSTAGE Ambassadors will be adept in the types of technology to be incorporated in DataSTAGE, e.g., command-line facile and distributed computing.

These Ambassadors may work very closely with the DataSTAGE teams serving a consultative function to accelerate scientific outcomes by HLBS investigators. Ambassadors should be willing and enthusiastic representatives for DataSTAGE and comfortable with learning from success and failure. For their time and help, Ambassadors may receive a combination of early access to the DataSTAGE platform, free compute time, monetary support for time, and relevant travel expenses will be covered (see the "User Engagement" section for more information).

Ambassadors will be required to complete specific small group training. DataSTAGE intends to integrate these trainings into existing professional events, if possible. The STAGECC will establish specific communication channels for these Ambassadors that will include specialized web pages and email announcements. Following the initial training and onboarding as Ambassadors, STAGECC will host a number of general interest trainings that will include using GitHub and data security. We will survey Ambassadors for input on training, as well as the platform, and will continually seek to revise materials and approaches.

According to interest, Ambassadors will be invited to become DataSTAGE trainers using the Carpentries' Instructor Training program. This workshop will help Ambassadors learn how to:

- Deliver lessons effectively
- Promote an inclusive atmosphere that adheres to the code of conduct
- Modify and update existing lessons using best practices in curriculum design
- Run assessment and evaluation for feedback to tool developers
- Provide feedback into the teaching program

In addition to covering relevant travel costs, Ambassador Trainers will receive a stipend in recognition of their extra effort.

Beta-User Training

Initial training will be focused on support for Ambassadors leveraging the DataSTAGE platform as part of a User Narrative. As additional users are onboarded, the training will broaden. Once the platform is available to a broader audience, we will support freely-accessible online training for Beta-Users at any time, as well as workshops led by the Carpentries-trained instructors, Ambassador Trainers, and DataSTAGE trainers. These instructors will also serve the role of providing feedback, either by communications or formal surveys, to the DataSTAGE developers on common issues encountered during training that can be included in the development backlog. Including a user feedback component within the training effort will further increase the touchpoints with potential users and serve to ultimately ground the platform development in the broader community needs.

An initial core workshop curriculum will include:

- DataSTAGE introduction
- Data responsible use and ethics training

- DataSTAGE onboarding
- Tracking problems with tutorials, walkthroughs, webinars, and other learning materials
- Becoming a DataSTAGE Ambassador

Useful documentation: Copper Team Training and User Engagement Plan

4.2 USER ENGAGEMENT

Along with plans to engage users through training, we recognize that DataSTAGE development teams will need to interact with specific users through the development of the Features that support a particular User Narrative. These users will be carefully identified in concert with the Steering Committee, NHLBI, and Data Stewards to ensure a handful of knowledgeable and appropriate users are virtually embedded within the development teams. This time commitment will necessitate compensation for time as consultants to the project.

The users will be provided with specific training and support documentation by the DataSTAGE teams so they can contribute to and test a User Narrative beginning at the earliest point of development. STAGECC will organize a centralized library of training materials and support documentation. Preliminary advertising of DataSTAGE to the general user will be through publications announcing the platform and its capabilities.

Community Outreach

The end goal of DataSTAGE community outreach is a sustainable and engaged community of scientific researchers who both use and invest in the DataSTAGE platform because of the utility it presents to their research. The DataSTAGE Consortium will seek to help DataSTAGE partners grow their own training programs that can expand upon existing training materials. The STAGECC will provide support to these efforts through assistance in materials development, video conferencing, and/or meeting hosting in an effort to establish users as training leaders, as well as facilitating integration of these materials into the training library.

4.3 ASSESSMENT

Successful operationalization of the Implementation Plan will require ongoing, but targeted, assessment. Assessment is directed towards gaining insight into alignment between the DataSTAGE Strategic Vision, the particular Work Streams driving implementation, and feedback from users. In-progress assessment related to the development of Epics and Work Plan activities is built into the Agile development process. Project management assessment and operational metrics are addressed in the Project Management Plan.

Training assessment includes activities such as:

- Online Pre-, Post-, and Follow-up Short Surveys

Ambassador and Beta-User assessment includes activities such as:

- Focus Group/Ethnographic analysis
- Online Pre-, Post-, and Follow-up Short Surveys

Community Outreach assessment includes activities such as:

- Tracking Consortium membership recruitment
- Usage patterns of the platform
- Identification of new collaborative projects

Training workshops will utilize Poll Everywhere or a similar tool to gather feedback on the training experience when the event closes. Additional surveys to gather later impressions will be circulated; the results synthesized and incorporated in the next round of training.

5 SUMMARY

Over the course of this Implementation Plan, we anticipate that there will be some evolution of the User Stories and Features as science and technology evolve. However, by 2021, we anticipate that the DataSTAGE platform will serve as a novel, fully-functioning resource in which users from a variety of disciplines and levels can perform complex operations and access newly-available scientific data to make significant strides in research and beyond.

6 REFERENCE DOCUMENTS

- Strategic Framework Plan
- Project Management Plan
- NHLBI DataSTAGE 60 Day o16n Plan v1-2 (drafted by the Operationalization Tiger Team)
- DataSTAGE User Narratives, Features, and Epics
- STAGE-RFC-2_DataSTAGE_Strategic_Planning_Nomenclature

REFERENCES

"A Strategic Framework for Data Storage, Toolspace, Access, and Analytics for biG-data Empowerment (DataSTAGE), v.1.0", DataSTAGE Coordination Center, February 2019. Strategic Framework

"Data Storage, Toolspace, Access, and Analytics for biG-data Empowerment (DataSTAGE) Project Management Plan, v.1.0", DataSTAGE Coordination Center, February 2019. Project Management Plan

"STAGE-RFC-2 DataSTAGE Strategic Planning Nomenclature", DataSTAGE Coordination Center, February 2019. STAGE-RFC-2_DataSTAGE_Strategic_Planning_Nomenclature

"DataSTAGE", Jonathan Kaltman, presented to Dr. Gibbons, NHLBI, Jan 22, 2019. Jon's presentation on Jan 22 to NHLBI/Dr. Gibbons

"DataSTAGE User Narratives, Features, and Epics", DataSTAGE Consortium Members, February 2019. DataSTAGE User Narratives, Features, and Epics

"Training and User Engagement Plan by Copper Team 2018", 5M.5. Product, DCPPC Copper Team (Titus Brown et al), 2018, Training and User Engagement Plan

"Progress Toward Cancer Data Ecosystems". Grossman, Robert L., PhD. The Cancer Journal: May/June 2018 - Volume 24 - Issue 3 - p 126–130 doi: 10.1097/PPO.0000000000000318

"STRIDES Initiative – NIH Common Fund." National Institutes of Health, U.S. Department of Health and Human Services, commonfund.nih.gov/strides.