A Strategic Framework for Data Storage, Toolspace, Access, and Analytics for biG-data Empowerment (DataSTAGE)

V1.0 - 20190426

DataSTAGE

A Strategic Framework for DataSTAGE

V1.0 - 20190426

Document Status

Version

V1.0

Approvals

Signatures presented below denote review and approval of the DataSTAGE Strategic Framework. These approvals are given based on the understanding that the Strategic Framework, 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+	03/28/2019

NIH Approvals:

Responsible Person	NIH NHLBI DataSTAGE Role	Approval Date
Jonathan Kaltman, NHLBI	Program Manager	04/26/2019
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
20191101	V0	N/A	Draft document created.
20190305	V0.1	Marcie Rathbun	In section 8, added link to Operationalization document: NHLBI DataSTAGE 60 Day o16n Plan v1-2
20190313	V0.2	Rebecca Boyles	User Narrative edits from consortia review incorporated
20190426	V1.0	NHLBI	V1.0 reviewed and approved by NHLBI Links & editing updates [Marcie]

TABLE OF CONTENTS

1 INTRODUCTION

1.1 PURPOSE OF THE STRATEGIC FRAMEWORK

The DataSTAGE Strategic Framework identifies the mission and vision of the DataSTAGE program and describes how the program will align across stakeholders to execute on common goals and how that performance will be measured. In the creation of this Framework and the complementary Implementation Plan, we focus the Consortium on a common goal, agree on actions, align resources, and prioritize needs.

2 EXECUTIVE SUMMARY

The purpose of the DataSTAGE Strategic Framework is to articulate a forward-looking path for the DataSTAGE Consortia and stakeholders to align across a complex Heart, Lung, Blood, and Sleep (HLBS) landscape of technologies, science, and data. The Framework is a culmination of an in-depth process that involved strategic analysis of the data, applicable methodologies, and needs with the key DataSTAGE stakeholders. This analysis was then developed further into the Framework document. The Framework is evergreen and will be regularly amended to reflect new priorities. The Framework was envisioned and created with guidance from NHLBI and the DataSTAGE Consortia.

The Strategic Framework consists of a mission, vision, and values, as well as overarching User Narratives and the orthogonal work streams that comprise the types of work needed to execute the DataSTAGE program.

A separate Implementation Plan, which maps goals, objectives, and strategies into specific Features, accompanies the Strategic Framework. The Implementation Plan, coupled with the Project Management Plan, establishes priorities, accountabilities, success indicators, and timeline and resources for projects. To create project priorities and transparency, the Implementation Plan uses the following guiding principles: availability of resources, impact on the NHLBI mission, return on investment, the utilization of technologies that maximize data security and integrity, and the implementation of cost-effective solutions.

3 BACKGROUND

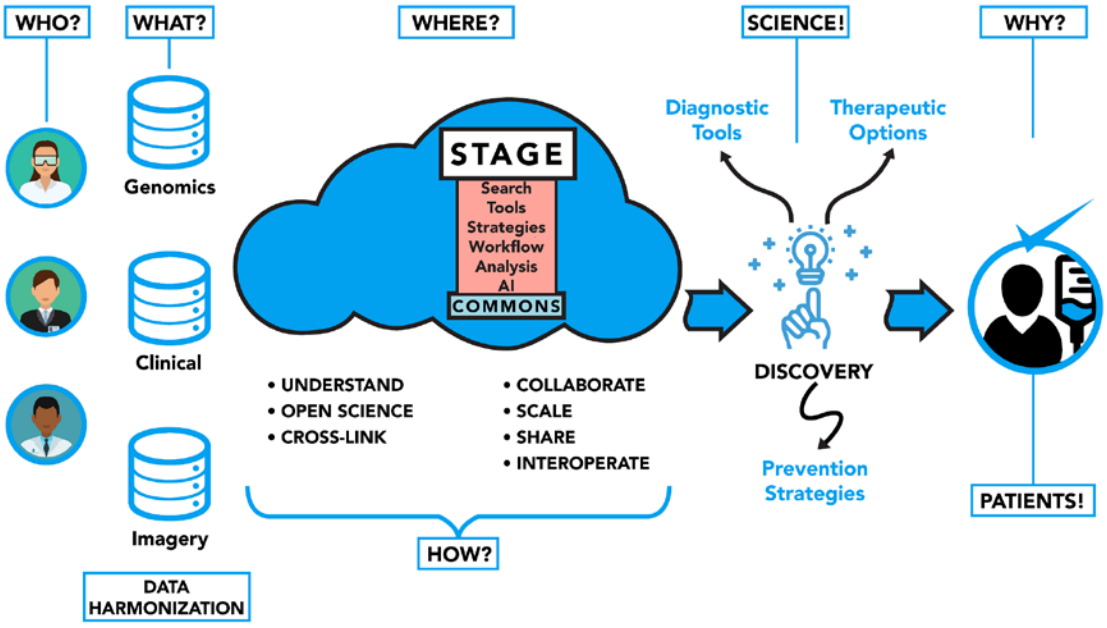
3.1 PROBLEM STATEMENT

Much has been written about the explosion of biomedical data that has been largely driven by the genomic revolution (Collins, Morgan, and Patrinos 2003; Green, Watson, and Collins 2015). In addition to the increasing availability and volume of genomic data, researchers and clinicians have seen a dramatic increase in data through the adoption of high-throughput assays. The need to leverage these data resources through the application of emerging data science approaches is recognized in the NHLBI Strategic Vision (National Heart, Lung, and Blood Institute, and Others 2016).

Modern HLBS research must now operate across a diverse data landscape that includes large data resources in high-throughput genomic, proteomic, metabolomic, microbiome, personal wearable, behavioral, and clinical domains. To support this work, advancements are needed in our ability to provide researchers with cost-effective and rigorous storage, management, tooling, and computation within their current workflows while upholding the NIH's responsibility to appropriately manage human subject data.

3.2 WHAT IS DATASTAGE?

In 2007, Jim Gray famously described a Fourth Paradigm of Science, in which science of the future would leverage interoperable knowledge and data online (Hey et al. 2009). The NHLBI DataSTAGE is an instance of a Data Commons, where HLBS researchers can go to find, search, access, share, store, crosslink, and compute on large scale data sets. It will be a cloud-based platform that has, at its foundation, a Commons that provides controlled access to data, tools, applications, and workflows to enable these capabilities in secure workspaces. DataSTAGE will accelerate efficient biomedical research and maximize community engagement, productivity and discovery.



4 MISSION

Critical to the success of the DataSTAGE program is consensus through a mission statement that articulates what DataSTAGE will provide for the user, developer, and programmatic communities.

The NHLBI DataSTAGE's *mission* is to develop and integrate advanced cyberinfrastructure, leading-edge tools, and FAIR data to support the NHLBI research community and accelerate discovery.

---

5 VISION

The DataSTAGE Consortium is jointly working towards a common future vision that will drive our implementation and management actions.

The *vision* for DataSTAGE is to be a community-driven ecosystem implementing data science solutions to democratize data and computational access to advance Heart, Lung, Blood, and Sleep science.

---

6 CONSORTIA VALUES

In all of our work as the DataSTAGE Consortia, we remain committed to a set of values that guides our thinking and ideas as an organization.

The DataSTAGE Consortia are committed to:

- Engagement with stakeholders to inform development;
- Responsible stewardship of NHLBI data assets and resources;
- Respect for the study participant and individual consent;
- Service to scientific advancements and HLBS health; and
- Alignment with the NHLBI Strategic Vision, NIH Data Science Strategic Plan, and related emerging data-intensive initiatives.

7 DESIGN PRINCIPLES

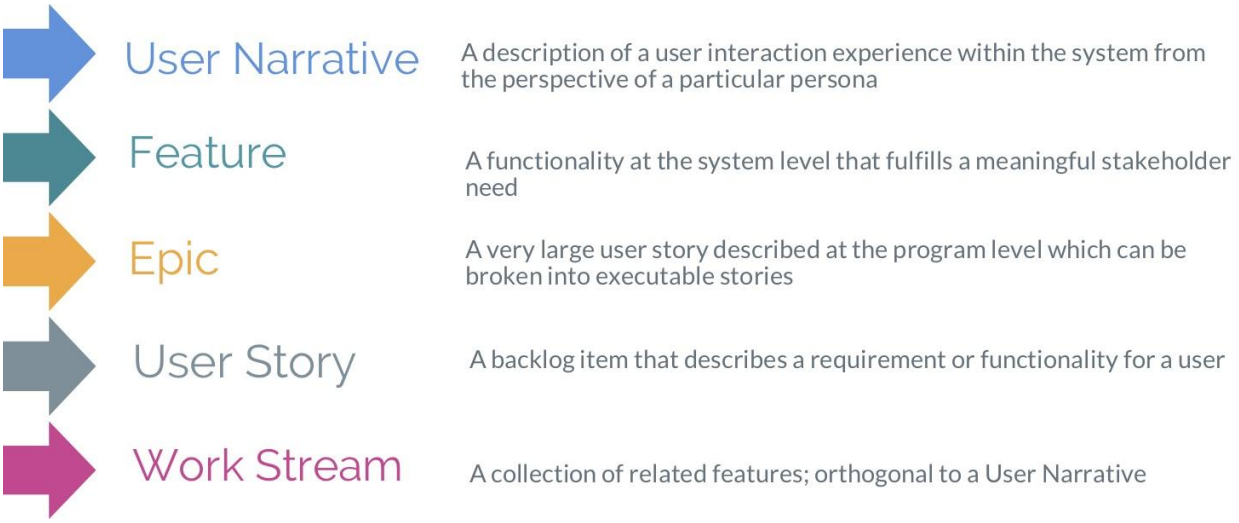
Design Principles are common guidelines or considerations that inform the approach to the DataSTAGE development. Here we highlight a number of high-level Design Principles that are cross-cutting across the User Narratives for DataSTAGE.

The cross-cutting Design Principles are:

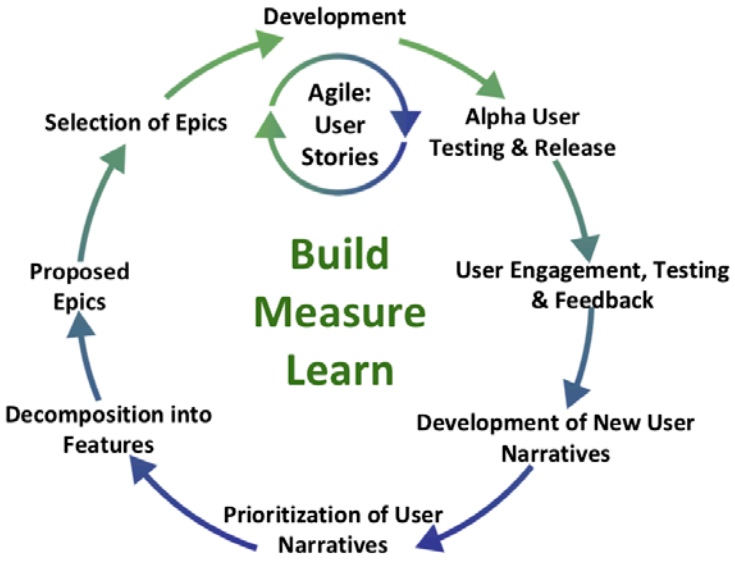
- Meet user needs and incorporate feedback
- Leverage existing tools and infrastructure, when feasible
- Do not duplicate infrastructure components
- Duplicate functionality when intentional and reasonable
- Architect interoperability with relevant systems
- Encounter a seamless experience, regardless of underlying components

- Leverage cost-advantageous cloud resources
- Support scalability and extension of functionality
- Have an early impact on computational-driven HLBS science
- Enable consistent, easy access to applications and tools for users across DataSTAGE
- Provide systems security for hosting identifiable data

We will use the vocabulary below to discuss the various levels of work breakdown for DataSTAGE.



These terms are drawn directly from the Agile literature in consultation with NHLBI, but many Agile methods use alternative terminology. Additional details can be found in the DataSTAGE Implementation Plan. Further, we are using a "Build Measure Learn" design cycle, as shown to the right. This provides us with a substantial degree of flexibility with respect to the User Stories, while maintaining the discipline and collective focus on the overall objective of DataSTAGE through less frequent modifications of the User Narratives.



7.1 USER NARRATIVES

One way to describe the intended outcome of the DataSTAGE program is through User Narratives. User Narratives are descriptions of a user interaction experience within the system from the perspective of a particular persona. Within the Implementation Plan, these User Narratives will further be broken into Features, Epics, and User Stories, as appropriate.

As is further described in the DataSTAGE Implementation Plan, stakeholder feedback and user testing will drive the prioritization and further development of DataSTAGE User Narratives. Overall this approach will provide a flexible, coordinated framework to drive DataSTAGE development while integrating user needs. Tasks necessary to accomplish User Narratives can also be organized into Work Streams, which are orthogonal to a User Narrative. Work Streams (see graphic below) group similar activities together to present an alternative view of progress towards the DataSTAGE vision. The below figure illustrates how our first User Narratives can map to broader Work Streams. This view also provides additional insight into how our User Narratives help the Consortia meet objectives.

Additional detail on the structure of the work hierarchy can be found in the DataSTAGE Implementation Plan.

	Production	Analysis Suite	Data Access	Data Management	Engagement
User Narrative: COPDGene Deep Learning **User Narrative:** Cohort Search **User Narrative:** TOPMed GWAS	Establish operationalized platform Implement security infrastructure Enable interoperability between Gen3, FireCloud, SevenBridges, and STRIDES providers	Deploy I2B2 & deep learning apps Geno/pheno search for all TOPMed datasets TOPMed GWAS Pipeline porting	Develop whitelist service with Gen3 Execute Data Access Authorization between NHLBI and developers Enable developer access to all TOPMed	Transfer COPDgene data & TOPMed genomic data to STRIDES cloud provider Support dataset updates within STRIDES cloud provider Transfer harmonized phenotype data from TOPMed DCC to STRIDES cloud provider	Onboarding of initial investigators User & training documentation Scientific publication announcing STAGE

Critically, User Narratives will be used to benchmark progress towards the DataSTAGE vision by asking potential users to complete a pre-specified narrative and documenting any shortcomings, trouble spots, and opportunities for improvement. These identified issues will be funneled into the program's development backlog.

User Narratives offer an opportunity to engage potential users in the development process, with regular feedback opportunities (e.g., sprint demos) to ensure that DataSTAGE is executing towards the vision, but also meeting future user's needs, even as they evolve over time. It is anticipated that new User Narratives will be refined through this process and will be incorporated into the Strategic Framework and Implementation Plan materials. DataSTAGE will remain connected to the user community, remain agile, and will intentionally evolve to drive future development efforts forward. User Narratives may represent near-term partial solutions that are deployed in stages to solicit user feedback and allow for rapid development.

The below are abbreviated User Narratives formulated in rough six-month timelines with more detail in the near term and less detail further out, with plans to refine as DataSTAGE progresses. The complete User Narratives will be maintained in the DataSTAGE User Narratives, Features, and Epics document and will be regularly reviewed and edited under a mechanism that is under development.

June 2019

1. a) A pre-approved group of computationally savvy researchers can create a user profile on the DataSTAGE environment using their eRA commons identity. b) They can access molecular and phenotypic data from selected TOPMed datasets for which they have approval and can use a simple cohort search to find and explore phenotypes using an interface that returns counts of study subjects within a single study matching search criteria. c) They can use a visual or programmatic (API) interface to perform computationally demanding (alignment/variant calling, etc.) batch analysis Jupyter notebooks or Rstudio to perform interactive analysis. d) They can collaborate on analyses (including sharing scripts, results, etc.) with other approved researchers.

2. a) A pre-approved group of computationally savvy users can test *proof of concept* functionality to understand and optimize cloud costs associated with running large scale computing or interactive analysis. b) As Alpha Users, they will be able to browse phenotypic and annotations of genomic variables within a single TOPMed study and view standard statistics on returned results.

December 2019

3. Researchers can access a DataSTAGE portal to register a profile associated with their ORCID or eRA Commons credentials and search the catalogue of DataSTAGE data and tool resources to identify datasets and tools relevant to their research needs.

4. Researchers with data approvals can use a hardened functionality described as proof of concept in User Narrative #2.

5. a) TOPMed and COPDGene researchers with familiarity with computational biology and approval can search and view all TOPMed molecular data (including pre-release data) and an expanded set of phenotypic data on the DataSTAGE platform. b) They can use visual and programmatic tools to explore, query, and analyze phenotypic and annotations of genomic variables across multiple studies based on a limited set of harmonized variables and obtain aggregate counts. c) They can upload and annotate private data and conduct combined analysis with DataSTAGE and TOPMed using docker-based (e.g., Deep Learning) notebooks and workflows.

6. a) To support the future integration of Cure Sickle Cell data for the Minimal Viable Product (MVP) system; architects and data managers must understand requirements for onboarding Cure Sickle Cell data and users. b) System architects have defined and documented procedures for on-boarding specialized tools into DataSTAGE for Cure Sickle Cell.

7. All users can access a minimal help desk and documentation through the DataSTAGE portal that addresses onboarding, data access, and common functionalities. A process for continuous improvement in training and materials, as well as potential changes to DataSTAGE based on user experience, is developed and provided to all users online.

June 2020

8. a) TOPMed approved researchers can log into DataSTAGE and access the DataSTAGE-hosted University of Michigan imputation server to upload their sparse genotype data via web browser or SFTP, convert it to VCF, and configure and run the imputation. b) The results are stored within a private DataSTAGE workspace for further analysis. They can receive notification of analysis status and use debugging tools to optimize or troubleshoot workflows.

9. a) A pre-approved group of harmonization experts can test proof of concept functionality to the data harmonization support interface to explore the phenotype data, choose which variables to use from different studies, decide on a set of transformations to apply to each variable, and then apply those transformations to create a combined, harmonized variable. b) Approved users will be able to test tools to support phenotypic data quality control; cleaning and harmonization and create synthetic cohorts by combining harmonized data across sets for later analysis.

10. a) Computationally savvy HLBS researchers without existing approval for TOPMed datasets can create accounts on DataSTAGE, browse across clinical variables in multiple studies, and view aggregate counts and matching sample criteria before requesting data access. They can search across studies and tools including Kids First and AnVIL resources. b) They can choose to upload private sparse genotyping data and use aggregated reference data to impute variants and then query across the Model Organism Databases for related results. c) They may request access to controlled access data (including data not governed by dbGaP) and receive approvals in an expeditious manner through DataSTAGE.

11. a) All DataSTAGE system users can utilize a hardened system with updated data releases, visualization tools for interactive exploration, workflow configuration tools, and phenotype harmonization tools. b) A user may interactively explore GWAS results from other population studies and create synthetic cohorts. c) Data access will be programmatically enforced. d) They can make use of a rich knowledge center of training and tutorials and participate in outreach events.

December 2020

12. TOPMed and COPDGene researchers with minimal computational biology experience can access computational biology support services (expanded help desk) to configure workflows and run analyses.

13. Qualified bioinformaticians can search and compute in the Cure Sickle Cell MVP using a command-line interface.

14. DataSTAGE system operators can use improved and more automated methods of

integration of new interactive applications to the DataSTAGE ecosystem, provide data access logs to guide data storage decisions and future data needs, and demonstrate enhanced processes in support of security and compliance considerations.

June 2021

15. a) All DataSTAGE system users can perform an expanded, harmonized search across TOPMed, AnVIL, Kids First, and Cancer Research Data Commons datasets to find participants relevant to their research question and work with publicly-available dbGaP data to perform searches to identify variables relevant to specific phenotype concepts. b) They can request data across synthetic cohorts and can publish analysis results according to appropriate compliance considerations for discovery and re-use.

16. Computationally savvy DataSTAGE researchers can conduct integrated analysis with data from other data ecosystems (Human Cell Atlas, HuMaP, All of Us, MVP. etc.).

December 2021

17. All DataSTAGE system users can access improved mechanisms for integrated analysis across other data ecosystems and access environmental data and appropriate analysis methods to interrogate environmental exposures.

18. System operators have documented the current system and plans have been made in coordination with NHLBI and the STAGECC for transfer of the resource to a sustainable location for maintenance and ongoing development.

8 REFERENCE DOCUMENTS

- Implementation 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

APPENDIX A: REFERENCES

Collins, Francis S., Michael Morgan, and Aristides Patrinos. 2003. "The Human Genome Project: Lessons from Large-Scale Biology." *Science* 300 (5617): 286–90.

Green, Eric D., James D. Watson, and Francis S. Collins. 2015. "Human Genome Project: Twenty-Five Years of Big Biology." *Nature* 526 (7571): 29–31.

Hey, Tony, Stewart Tansley, Kristin M. Tolle, and Others. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Vol. 1. Microsoft research Redmond, WA.

National Heart, Lung, and Blood Institute, and Others. 2016. "Charting the Future Together: The NHLBI Strategic Vision." *Bethesda, MD: NHLBI*.