



# Northeast Big Data Innovation Hub Strategic Plan

NSF Grant [1916585](#)

June 1, 2020

For Public Distribution

PI - Jeannette Wing / Columbia University

Co-PI - Andrew McCallum / University of Massachusetts Amherst

Co-PI - James Hendler / Rensselaer Polytechnic Institute

Co-PI - Vasant Honavar / Pennsylvania State University

Executive Director - Florence Hudson / Columbia University

## Table of Contents

1. Purpose - Mission, Goals, Focus Areas, Objectives, Project Activities	3
2. Community Engagement	8
3. Management and Staffing Plan	9
3.1 Project Team	9
3.2 Steering Committee	10
3.3 Advisory Board	11
3.4 Seed Fund Steering Committee	11
3.5 Project Activity Leaders	11
3.6 National Coordination Committee	12
4. Project Governance	12

# 1. Purpose - Mission, Goals, Focus Areas, Objectives

## 1.1 Mission Statement

The mission of the Northeast Big Data Innovation Hub (Northeast Hub) is to build and strengthen partnerships across industry, academia, nonprofits, and government to address societal and scientific challenges, spur economic development, and accelerate innovation in the national big data ecosystem.

The Northeast Hub is a community convener, collaboration hub, and catalyst for data science innovation in the Northeast Region. The Hub amplifies successes of the community, and shares credit across the community to encourage collaboration and mutual success in data science endeavors.

The Northeast Hub region includes the states of Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island and Vermont.

## 1.2 Goals

The goals of the Northeast Hub are to:

1. Build collaborations to address real-world challenges through translational data science approaches
2. Foster innovation and scale endeavors that reflect regional interests and align with national priorities related to data science
3. Support and promote representative community engagement/impact across all Hub activities
4. Increase data science capacity and talent, emphasizing underserved communities

The Northeast Hub emphasis is on translational data science, working with the community to develop and leverage data science approaches to catalyze action from knowledge and insight, enabling the development of new solutions to societal problems through community-led efforts and innovations.

### 1.3 Focus Areas

In 2019, the Northeast Hub’s Priority Areas included four thematic areas and four cross-cutting areas (Figure 1), developed to address the goals listed above in collaboration with cross-sector communities across the Northeast. These priority areas are subject to change or modification based on community feedback and ongoing assessment of the value they add.

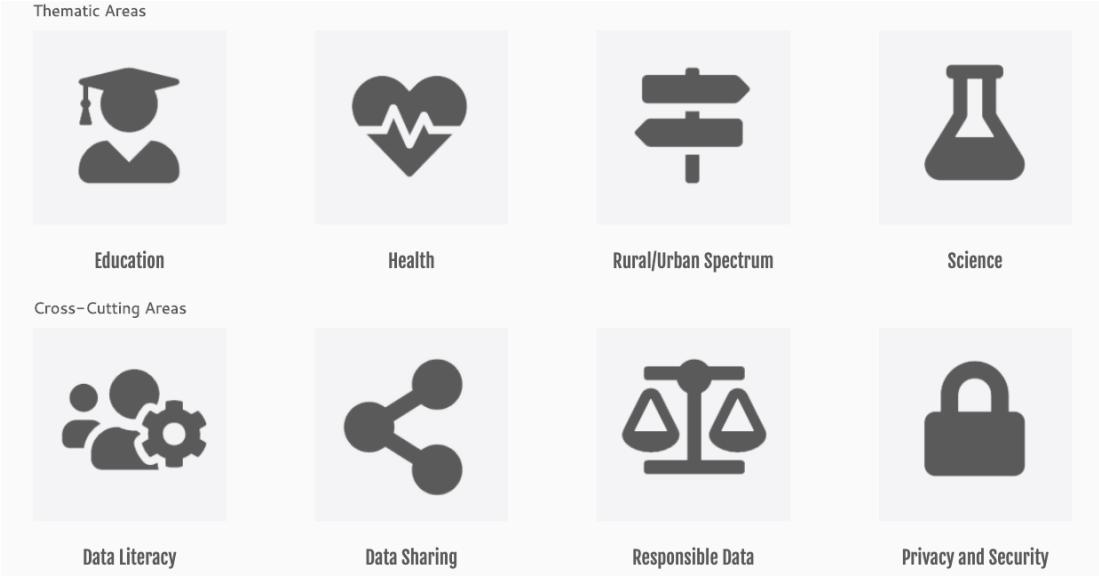


Figure 1: Northeast Hub 2019 Priority Areas - Thematic and Cross-Cutting

In March 2020, a reassessment of the eight priority areas was conducted by the Northeast Hub leadership team including the PI, Interim ED, Steering Committee and Advisory Board, resulting in the streamlining of the eight priority areas to four focus areas (Figure 2). These four focus areas were chosen to build on the Hub’s current strengths, activities, and regional distinctions. Streamlining to four focus areas enables the Northeast Hub team to filter project activities and align resources to enable progress and maximize value and positive impact for the community.



Figure 2: Northeast Big Data Innovation Hub Focus areas

## 1.4 Objectives

The Hub's four focus areas guide the work of the Hub, and of ongoing decisions made to create and support project activities to enable translational data science innovation and application to societal and scientific challenges. The objectives in each of the four focus areas are as follows.

**Health:** The objective of this focus area is to help advance data sharing, acquisition, integration, analysis and resulting insights in the pursuit of improved public health and health outcomes. This includes:

1. Enabling data sharing, acquisition and integration to develop knowledge and insight in the pursuit of improved health outcomes, including alternative health data sources such as environmental factors, social media and mobile health
2. Supporting the development and deployment of advanced analytics, including causal discovery and reasoning, artificial intelligence, and machine learning in biomedicine
3. Enabling data science collaborations in the support of precision medicine, including advanced decision support leveraging data to deliver customized and personalized knowledge, insight and recommendations

Efforts in this application area will include designing projects such as the Exposome Data Insight Pilot, and the COVID Information Commons. In response to the COVID-19 pandemic and the epicenter in the Northeast region, the Hub is increasing collaboration regarding COVID-19 research and community building, both across the Northeast and in partnership with the Midwest, South and West Hubs across the country. This includes a COVID-19 Resource webpage <http://nebigdatahub.org/covid19/> launched in March 2020, which is continually enhanced with COVID-19 resource links including potential funding sources, current NSF RAPID COVID awards, pertinent datasets, virtual events and collaboration opportunities in the U.S and internationally. This will be further developed into a COVID Information Commons in collaboration with the Midwest, South and West Hubs through NSF RAPID COVID award #2028999.

**Urban to Rural Communities:** The objective of this focus area is to facilitate activities and collaborations focused on improving the understanding of data ecosystems in urban and rural environments, and the potential application of data analytics and responsible data science practices to benefit rural and urban communities, citizens, and the environment. This includes:

1. Enabling acquisition and leverage of interoperable and accessible data sets for use across rural and urban communities, for data driven community and environmental planning

2. In collaboration with the Health focus area, determining the impact of urbanicity on health and other measures of well-being, for humans and their environment
3. Creating data science innovations to envision and enable sustainable environments, including collaborations incorporating data science, environmental science, and biogeochemistry

**Responsible Data Science - including security + privacy + ethics:** The objective of this focus area is to develop recommendations and best practices for Responsible Data Science (RDS) by design, with security, privacy and ethical use of data and data science methods. This includes:

1. Enabling the development and dissemination of responsible data science best practices, such as building and deploying of integrative data equity systems, incorporating ethical and legal norms in all stages of the data science life cycle, with security, privacy and ethics
2. Engaging IEEE standards development activities for the TIPPSS framework - Trust, Identity, Privacy, Protection, Safety and Security - and its use in clinical Internet of Things, connected healthcare, and smart and connected communities
3. Collaborating with cybersecurity researchers and domain scientists to determine their needs and challenges regarding trustworthy data, and data driven cybersecurity, privacy, and ethics

**Education + Data Literacy:** The objective of this focus area is to work with the community to guide the development and dissemination of data science approaches and resources for education across the community, from PreK-12 to higher education and industry, with the goal to improve data literacy and educational outcomes. This includes:

1. Amplifying the success and reach of data science education and data literacy activities, from virtual and in-person workshops to Massive Open Online Courses (MOOCs), to empower educators and researchers to improve data science practices and reach a broad community
2. Enabling community and educational institutions across the region to deliver data science education to underserved constituencies, extending the success of the Data Science for All workshops, and perhaps adding domain-specific data science and data carpentry workshops
3. Increasing community collaboration to leverage the Big Data for Education spoke best practices, including innovations in intelligent tutor pedagogy, and collaborative analyses of Big Data for Education MOOC data

## 1.5 Alignment of Goals, Focus Areas and Project Activities

The Northeast Hub goals provide the overarching guidelines for the focus areas and project activities as depicted in Figure 3. The project activities portfolio evolves over time, with the addition of new project activities developed with the community, evolution of existing projects, and quiescing of completed projects. You may contact the Hub to engage in or add your project to the Hub activities at [contact@nebigdatahub.org](mailto:contact@nebigdatahub.org).

New project activities may be added which align with the Focus Areas, based on Hub and community resources available and committed to achieve the objectives of the project.

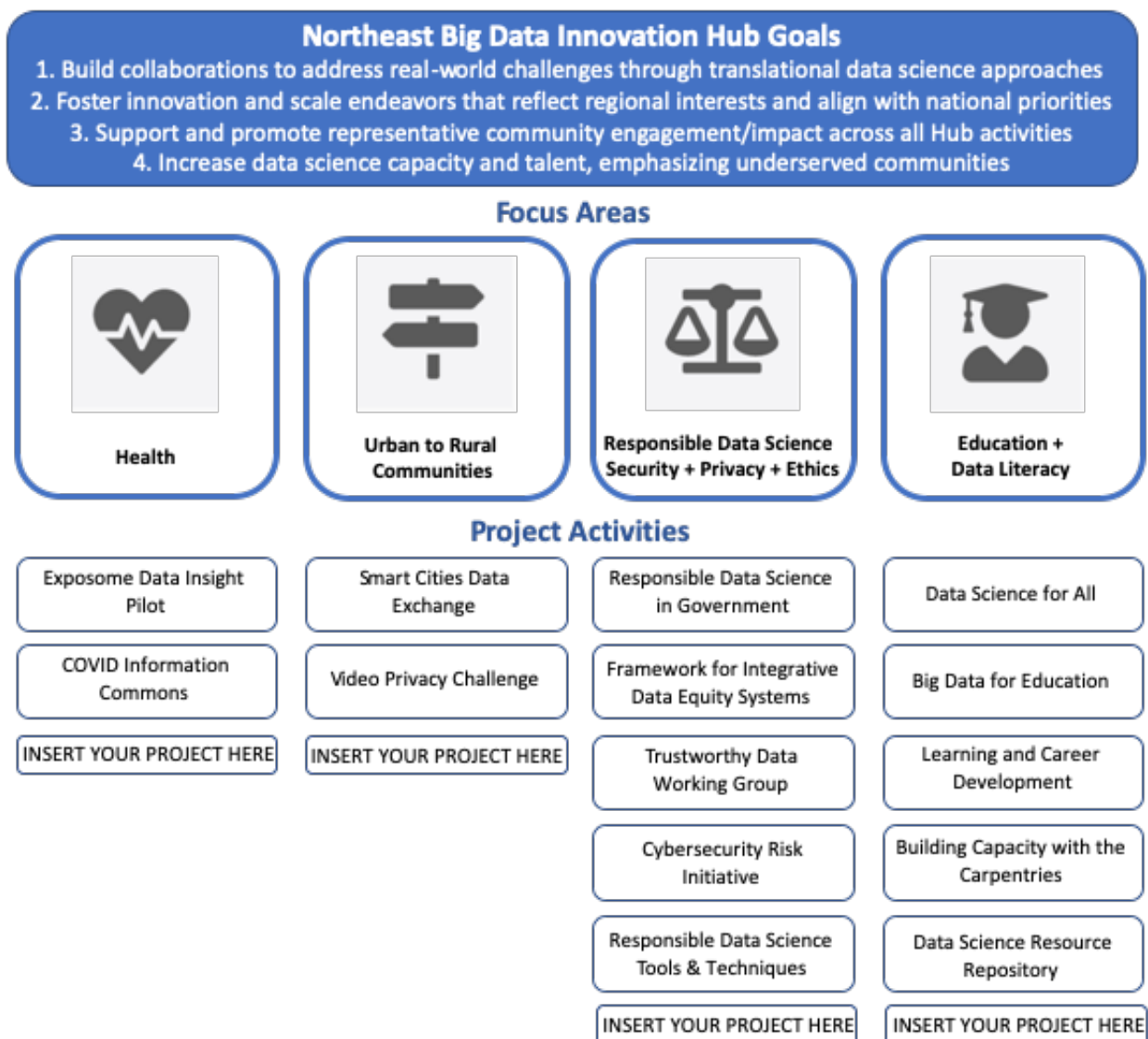


Figure 3: Northeast Big Data Innovation Hub Goals, Focus Areas, Project Activities

## 2. Community Engagement

The Northeast Hub is a community convener, collaboration hub, and catalyst for data science innovation in the Northeast Region. The Northeast Hub is increasing researcher and institutional outreach and communication across the community to identify research activities addressing real-world challenges through translational data science approaches, and to create new collaborations in this endeavor. The Hub Steering Committee and Advisory Board are leaders in enabling community engagement, including identifying new projects of interest. For instance, one of the Hub Co-PIs and along with one of the Advisory Board members are Project Activity Leaders (PALs) on the Exposome Data Insight Project in the Health focus area.

The Hub is increasing outreach to Data Science Institutes (DSIs), Vice Provosts of Research (VPRs) and their counterparts in smaller institutions, to identify and amplify data science related collaboration opportunities and needs across the Northeast. This includes identifying potential pilot projects for consideration by the Seed Fund Steering Committee (SFSC). The SFSC was created in May 2020 to provide pilot funding for new projects which align with the Hub focus areas, as explained in the Section 3 Management and Staffing Plan. The SFSC will work with the Executive Director's Office to engage the Northeast Hub communities broadly through VPRs, DSIs, collaborators, and various communications channels to reach new collaborators, extending our reach beyond the 600+ individuals who have already engaged with the Hub from more than 45 collaborator institutions <https://nebigdatahub.org/collaborators/>. To add your institution as a collaborator, please contact the Hub via email at [contact@nebigdatahub.org](mailto:contact@nebigdatahub.org).

Outreach activities through events such as the NSF Harnessing the Data Revolution (HDR) Principal Investigators (PI) meeting have identified new collaboration opportunities across the Northeast Hub region, and across the country, providing the opportunity to build on current project activities to further the research in addressing real world challenges. One such example is the identification of HDR PIs who have been introduced to the Exposome Data Insight Project PALs to enable collaboration focused on improving population health insights and potentially healthcare outcomes through genomic, phenotypic, and environmental data analytics.

The Hub community communications plan is executed through multiple virtual channels including distribution of a monthly newsletter with timely and pertinent information regarding the Hub focus areas, project activities, collaboration opportunities, and community events. The Hub uses social media channels such as Twitter, YouTube and LinkedIn to disseminate timely information, original content, and best practices including reports and other publications. Community members can sign up to be added to the Hub newsletter and view the social media resources at the Hub website <http://nebigdatahub.org/connect/>. The website is updated dynamically with new information of use to the community, such as the addition of a COVID-19



Resources webpage in March 2020 <http://nebigdatahub.org/covid19/> to provide links to research funding opportunities, grants awarded, datasets, and collaboration opportunities.

The Hub is adding Success Stories to the website and communications plan in 2020, highlighting community activities which have accomplished significant project goals. This will include information regarding project accomplishments, manuscripts, papers and workshop proceedings in the Hub’s four focus areas, as well as requests to enable and further collaborative research partnerships stemming from these project and activity success stories.

### 3. Management and Staffing Plan

The Northeast Hub is a community of collaborators aligned to the Hub’s goals, focus areas and objectives as described in sections 1 and 2 of this strategic plan. There are six overall entities in the management and staffing of the Hub organization as shown in Figure 4 and outlined below.

#### Northeast Big Data Innovation Hub Organization (\* denotes Volunteers)



Figure 4: Northeast Big Data Innovation Hub Organization Chart

#### 3.1 Project Team

The Project Team consists of the Principal Investigator, Co-Principal Investigators, and the Executive Director’s office.

The Principal Investigator (PI) serves on the Hub’s Steering Committee as chair and permanent voting member. The Co-Principal Investigators (Co-PIs) are regional leaders whose insights help further the Hub’s mission and who have demonstrated sustained interest to date in engaging with the Hub in a voluntary role. The co-Principal Investigators may serve on the Hub’s various committees according to the policies and procedures set forth for each respective committee, as Project Activity Leaders, and may otherwise be convened or consulted on an ad hoc basis to provide input on specific projects and activities of the Hub based on their areas of expertise.

The Executive Director’s office (EDO) has responsibility for the operations of the Hub in a manner that is responsive to the needs of the Hub community. The EDO office will also manage and administer the Seed Fund of \$250,000 annually working with the Seed Fund Steering Committee, under the Hub Steering Committee’s guidance. Full-time staffing of the Hub includes an Executive Director (ED) and Operations Manager (OM). The Executive Director will lead strategy development and execution, including working with the Steering Committee to refine the Hub’s strategic direction; overseeing the execution of Hub projects and activities; and conceiving, designing, and supporting partnerships. The Executive Director is an ex officio non-voting member of the Steering Committee and Seed Fund Steering Committee. The Operations Manager is responsible for Hub operations, including planning and executing special events and meetings, and leading the strategic communications project for the Northeast Hub and, as part of the national coordination effort, for all Hubs.

### **3.2 Steering Committee**

The Steering Committee (SC) provides oversight for the Hub. Its membership reflects the diversity of the Northeast Hub community across backgrounds, sectors, and geographies, including underrepresented and underserved communities. The Steering Committee shall be no more than seven members. The Hub has revised its existing bylaws to define selection and approval processes that ensure membership is refreshed on a rotating basis. The SC and Project Team nominate new SC members. The SC selects new SC members from the nominees via an internal vote. The Executive Director (ED) will serve as a permanent ex officio member of the Steering Committee. The Steering Committee members will convene quarterly and are expected to play an active role in advancing the mission of the Hub, to advise and assist the Executive Director in the management of the Hub, and to broadly represent the interests of Hub stakeholders.

### **3.3 Advisory Board**

The Advisory Board (AB) consists of community members with expertise in the Hub's focus areas. They will provide ongoing advisory support for the focus area activities and include Project Activity Leaders (PALs). The Advisory Board may be convened or consulted on an ad hoc basis to provide non-binding input on specific Hub projects and activities based on their areas of expertise, and to discuss any updates to activities based on evolution of the Hub plan and input from the community. Advisory Board membership approvals will be made on a regular rolling basis each calendar year to ensure an infusion of new perspectives. Advisory Board members are nominated by and selected by the Steering Committee.

### **3.4 Seed Fund Steering Committee**

The Seed Fund Steering Committee (SFSC) will administer the seed fund and define the policies and processes for allocating the \$250,000 available annually for seed funding for the benefit of competitively selected projects. The SFSC was initially composed of the members of the Hub Steering Committee. The inaugural Chair of the SFSC was selected by the Northeast Hub Steering Committee and appointed May 1, 2020. The SFSC Chair nominates a slate of SFSC members to the SC who approve the nominees. The SFSC Chair contacts the approved nominees to recruit them to join the SFSC. The inaugural SFSC will be in place by August 1, 2020, with no more than one member of the BD Hub's Project Team as a voting member, and 3-5 additional members selected by the SFSC Chair and approved by the SC, demonstrating membership diversity. The SFSC will develop the Seed Fund process, including identification, assessment and funding of pilot projects. Seed funding may be provided for pilot activities such as workshops, webinars, travel, and student engagement. The Seed Fund Steering Committee will lead the development of the Seed Fund process with the SFSC in collaboration with the Hub Executive Director and Steering Committee. The Seed Fund pilot funding may be complementary to funding from industry, government, academia and non-profits for the most promising projects.

### **3.5 Project Activity Leaders**

The Project Activity Leaders (PALs) are Northeast Hub community members leading specific project activities aligned with the defined focus areas. The PALs may include collaborators from the Midwest, South and West Hubs. The PALs work with the ED to define project objectives, identify collaborators, and engage the broader Hub community as appropriate to meet the objectives of the project.

### **3.6 National Coordination Committee**

The National Coordination Committee (NCC) is a small, nimble body that allows autonomy of the regional Hubs while facilitating grassroots coordination among the four regions. Each Hub's Executive Director will serve on the NCC, plus at least one additional member from each region, with each Hub having an equal number of representatives. The Hub's representative will be selected by the Steering Committee from the membership of the Steering Committee or the Advisory Board based on their overall knowledge of the Northeast Hub strategic priorities and strengths and their willingness to represent these for the benefit of the broader Hubs' ecosystem. One member of each Hub's leadership team will serve as the NCC's point of contact with the National Science Foundation as assigned on a rotating basis of one year among all of the Hubs.

## **4. Project Governance**

Project activities are led by Project Activity Leaders (PALs) who are community volunteers, including co-PIs, Steering Committee and Advisory Board members, and other leaders from the Northeast Region and collaborators from other regions. For instance, the Exposome Data Insight PALs include a co-PI, an Advisory Board member, and the co-founder of an organization headquartered in the West Hub region with access to sensor testbeds across the country.

The Steering Committee and Advisory Board members are leveraged as thought leaders and PALs for project activities in the focus areas of their expertise. As we further develop project activities in the Hub Focus Areas, these SC and AB members provide guidance regarding projects to create, collaborators to engage, and objectives for the projects.

The Focus Areas are used as a guide and filter to prioritize opportunities and projects which have the potential to best leverage the strengths and resources of the Hub, and bring the most value to the Northeast Hub community. The Focus Areas are used to assess current projects, and to prioritize potential new projects including those brought forward by the community.

Community members are encouraged to bring forward existing and new projects aligned with the Hub's focus areas which may be leveraged and amplified with the Hub, and to engage in collaboration to further their research. Projects and activities which do not align with the current focus areas will not be brought into the Hub project portfolio.

The Hub operating model is based on leverage of the two full time staff members in the Executive Director's office, the ED and the Operations Manager, along with some part time resources and student interns, and volunteers across the Hub community. Due to the limited resources in the Hub staff, when new projects are brought forward to be included in the Hub portfolio we may have to adjust resources across the portfolio of Hub activities. This could precipitate re-assessment of the Hub project portfolio, with potential quiescing of projects which are assessed to be completed or not providing sufficient value to the Hub community.

Project activities will be assessed on an annual basis. This activity will be led by the Executive Director in collaboration with the leadership team consisting of the PI, EDO, SC, AB and SFSC as appropriate to determine project alignment with the Hub focus areas, and potential community value. Existing Hub projects will be assessed from three perspectives:

- 1) Affinity to the defined Hub focus areas
- 2) Progress made including activities and accomplishments to date
- 3) Hub and community resources available and committed to the project going forward

Based on this assessment, some projects may be quiesced. This could occur if the project is deemed not to align with the Hub focus areas, if the project or project team has completed their activities, if the project has not gained sufficient traction in the community, or if there is insufficient Hub or community resource available to commit to the project.

Completed projects will be transitioned to Success Stories when feasible, highlighting accomplishments and learnings from the project. The Hub project team will collaborate with the project leader to request a short Success Story which could be highlighted on the Hub website, and which may link to material results of the project such as papers, workshop materials, websites and collaboration opportunities related to the project.

There are three types of success stories which may be submitted to be published on the Hub website. The first are Success Stories communicating the project completion and success of a project, workshop, or research initiative in one of the focus areas, highlighting the value or insights delivered. The second is a Success Story which includes objects which can be leveraged for broader impact, such as workshop proceedings and agendas which will enable other workshops modeled after them to be developed to broaden the impact of the effort. The third type of Success Story includes a collaboration request, for instance requesting that researchers and students interested in collaborative analysis of big data for education contact one of the Success Story project leaders which could perhaps lead to new collaboration, insights and publications. All these types of Success Stories may be published on the Hub website. The Hub

may also enable further collaboration for the Success Story projects, including one-on-one connections and workshops or webinars to further the research and outcomes to extend collaborative success.

When a project has been assessed to be quiesced, the Hub will contact the project activity leader for a project learnings debrief to determine lessons learned. These lessons learned will be documented internally and shared with the Steering Committee.