Cyberinfrastructure Center of Excellence Pilot: Connecting Large Facilities Cyberinfrastructure

Ewa Deelman*, Anirban Mandal†, Valerio Pascucci§, Susan Sons‡, Jane Wyngaard¶
Charles F Vardeman II¶, Steve Petruzza§, Ilya Baldin†, Laura Christopherson†, Ryan Mitchell*, Loïc Pottier*
Mats Rynge*, Erik Scott†, Karan Vahi*, Marina Kogan‖, Jasmine A Mann*,
Tom Gulbransen**, Daniel Allen**, David Barlow**, Santiago Bonarrigo**,
Chris Clark**, Leslie Goldman**, Tristan Goulden**,
Phil Harvey**, David Hulsander**, Steve Jacobs**, Christine Laney**, Ivan Lobo-Padilla**,
Jeremy Sampson**, John Staarmann**, Steve Stone**

*Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA
{deelman, rmitchel, lpottier, rynge, vahi, jmann}@isi.edu
†Renaissance Computing Institute (RENCI), University of North Carolina at Chapel Hill, NC, USA
{anirban, ibaldin, laura, escott}@renci.org
‡Center for Applied Cybersecurity Research, Indiana University, Bloomington, IN, USA
sesons@iu.edu
§Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA
{spetruzza, pascucci}@sci.utah.edu
¶Center for Research Computing, University of Notre Dame, Notre Dame, IN, USA
{jwyngaar, cvardema}@nd.edu
‖School of Computing, University of Utah, Salt Lake City, UT, USA
kogan@cs.utah.edu
**National Ecological Observatory Network, Battelle Ecology, Inc., Boulder, CO, USA
gulbransen@battelle.org, {dallen, dbarlow, sbonarrigo, clarkcp, lgoldman, tgoulden, pharvey, dhulslander, sjacobs,
claney, ipadilla, jsampson, staarmann, sstone}@battelleecology.org

***Abstract*—The National Science Foundation's Large Facilities are major, multi-user research facilities that operate and manage sophisticated and diverse research instruments and platforms (e.g., large telescopes, interferometers, distributed sensor arrays) that serve a variety of scientific disciplines, from astronomy and physics to geology and biology and beyond. Large Facilities are increasingly dependent on advanced cyberinfrastructure (i.e., computing, data, and software systems; networking; and associated human capital) to enable the broad delivery and analysis of facility-generated data. These cyberinfrastructure tools enable scientists and the public to gain new insights into fundamental questions about the structure and history of the universe, the world we live in today, and how our environment may change in the coming decades. This paper describes a pilot project that aims to develop a model for a Cyberinfrastructure Center of Excellence (CI CoE) that facilitates community building and knowledge sharing, and that disseminates and applies best practices and innovative solutions for facility CI.**

***Index Terms*—cyberinfrastructure, large facilities, major research infrastructure, NEON**

I. Introduction

The National Science Foundation (NSF) and other governmental agencies in the United States have invested significant resources in the development of Large Facilities (LFs), recently also referred to as Major Research Infrastructure projects) that are at the forefront of scientific research and innovation. At the core of the LFs is cyberinfrastructure (CI) that manages instruments, data, and computing. Broadly, CI "consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible." [1].

Although there are functional commonalities between the CI of various LFs, a recent survey conducted by the NSF Cybersecurity Center of Excellence [2] found that all of the survey's LF respondents–15 in total–are individually and independently developing in-house software to fulfill their CI needs. They are sharing their experiences infrequently during workshops, such as the NSF's annual Cybersecurity Summit [3] or a series of NSF workshops on Large Facilities Cyberinfrastructure [1].

In 2017, the Large Facilities Cyberinfrastructure Workshop (LF CI Workshop) brought together community leaders to discuss the growing need of the LF to be part of a community that shares the most advanced cyberinfrastructure technology [1]. The workshop found that enabling synergistic interactions across the LFs and CI communities would be beneficial, as it would allow achieving sustainable development of a CI addressing the needs of current and future LFs. In particular,

978-1-7281-2451-3/19/$31.00 ©2019 IEEE
DOI 10.1109/eScience.2019.00058

the workshop found that "the need for, and benefits of, close interactions, collaborations, and sharing among the facilities and with the CI communities are well recognized, including the sharing of CI related expertise, technical solutions, best practices, and innovations across NSF Large Facilities as well as research facilities outside NSF (DOE, NIH, NASA, etc.)."

However, another finding of the workshop was that there is a "lack of effective mechanisms and funding structures to support interactions and sharing among facilities regarding their CI," and that "there is a critical lack of a focused entity that could facilitate interactions and sharing across facilities and a CI-centered community". As a result, the workshop recommended the establishment of "a center of excellence (following a model similar to the NSF-funded Center for Trustworthy Scientific Cyberinfrastructure, CTSC/Trusted CI [4]) as a resource providing expertise in CI technologies and best practices related to large-scale facilities as they conceptualize, start up, and operate."

This paper describes an effort to pilot a Cyberinfrastructure Center of Excellence [5] that directly addresses the community recommendations put forth by the 2017 LF CI Workshop and develops a plan for a CI Center of Excellence (CI CoE). The goal for the CI CoE is to serve the CI needs of LFs and large CI projects by:

1) building a community centered around CI for NSF Large Facilities;
2) creating a community-curated portal and knowledge base for the sharing of "CI-related challenges, technical solutions, innovations, best practices, personnel needs" [1]; and
3) defining an overarching entity for LFs "that can strategically address workforce development, training, retention, career paths, and diversity, as well as the overall career paths for CI-related personnel" [1].

The CI CoE Pilot project includes five academic institutions: the University of Southern California (project lead), the Renaissance Computing Institute at the University of North Carolina – Chapel Hill, the University of Notre Dame, the University of Utah, and Indiana University. The Pilot effort was funded by the National Science Foundation in the Fall of 2018 and is projected to last two years. The goal of the pilot is to develop a model and a blueprint for a CI CoE that facilitates community building and sharing, and applies knowledge of effective practices and innovative solutions to facility cyberinfrastructure. This paper describes the CI CoE Pilot's experiences and accomplishments during the first year of the project.

II. Piloting a Cyberinfrastructure Center of Excellence

As of June, 2019, there were two dozen LFs [6] that develop and operate sophisticated instruments in order to serve the scientific community in a variety of domains. LFs have constructed telescopes [7], [8], neutrino detectors [9], particle colliders [10], [11], ocean-going vessels [12], gravitational-wave detectors [13], ocean-cabled arrays [14], [15], and sophisticated towers capturing environmental data [16], among many other advanced instruments.

To better understand the specific CI challenges faced by LFs, the opportunities for cross-facility interactions, and the potential for long-term knowledge and capability building, the Pilot identified the National Ecological Observatory Network (NEON) [17] as the first LF with which to engage. As the Pilot was getting underway, NEON was working on improving their CI and was receptive to a potential collaboration with our project.

NEON is an ecological observation facility that collects and provides open data about the changes in North America's ecosystems. NEON's capture, processing, and dissemination of ecological data improves our understanding of our environment and provides more accurate forecasting of how human activities impact ecology [16]. NEON builds and operates various ecological sensors at a number of geographic sites in order to collect a rich set of data. These collection sites are strategically located across 20 ecoclimatic domains within the U.S. and represent regions of distinct landforms, vegetation, climate, and ecosystem dynamics. NEON sites are classified as either "Core" sites or "Relocatable" sites. Each ecoclimatic domain consists of two different types of sites: (1) a Core terrestrial site that collects data to characterize terrestrial plants, animals, soil, and the atmosphere, and (2) a Core aquatic site that collects data to characterize aquatic organisms, sediment and water chemistry, morphology, and hydrology. These Core sites are set up to collect data at the same location for 30 years and are designed to statistically capture and illustrate terrestrial and aquatic wildland conditions. Additionally, NEON has 27 Relocatable terrestrial sites and 13 Relocatable aquatic sites that are distributed throughout the ecoclimatic domains, as well. Data collection is standardized across all sites–Core and Relocatable, terrestrial and aquatic–and occurs at various spatial and temporal scales. Where logistically possible, terrestrial and aquatic sites are co-located to capture connections across atmospheric, terrestrial, and aquatic ecosystems. Automated instruments, observational sampling, and airborne remote sensing methods are used to capture and gather the data. NEON has standardized and integrated these collection methods to ensure the comparability of ecological patterns and processes between NEON sites through time.

In 2018, NEON was transitioning out of the construction phase and into the operation phase. It was also planning to conduct a number of enhancements to their CI. The CI CoE Pilot took this opportunity to propose an engagement activity with NEON to understand its objectives, learn its current practices (including both successful and those needing improvement), identify and provide technical expertise on state-of-the-art CI tools and methodologies that could be applied in the NEON environment, and distill and disseminate lessons learned that were of potential value to other LFs and the CI community.

There are a number of challenges when interacting with a Large Facility: the LF often has firm and often tight timelines for deliverables, it has well-established practices, which may

or may not be open to outside collaborations, and it has a clear mission focus that drives the projects to prioritize data and service delivery to their users.

To overcome these challenges, the Pilot worked closely with NEON to identify areas of potential engagement that aligned with the NEON enhancement timeline, that were of interest to both groups, and in which the Pilot could provide the required expertise and resources. The following were identified as potential areas of collaboration: web presence improvements, prototyping of new sensor gateways, exploring disaster recovery options, and prototyping new data management, data analytics, and data processing pipelines and workflows. To ensure the efficiency and success of this engagement, the Pilot identified and assembled the necessary expertise and dedicated the required effort to collaborate with NEON in a hands-on fashion. Some activities involved assisting with the evaluation of existing CI capabilities, and some included prototyping new CI solutions. As a result of the engagement, NEON has deployed some of the Pilot's suggested enhancements into their test infrastructure and is currently evaluating these upgrades for suitability in their production environment. The suggested identity management solution (described in Section IV-F) has been deployed on the main NEON website.

III. Engagement with NEON

To formalize the engagement process, the CI CoE Pilot developed an engagement model to employ with a single large facility. Fig. 1 illustrates this model. The model that takes an iterative, cyclical approach. Details for each step are provided below:

1) Engage with the LF, continuously and regularly interact with it to understand the goal of its CI enhancements and target community, and provide hands-on help and consulting. During the engagement, the Pilot and LF should strive to have both in-person meetings (at least 2 per year) and regularly scheduled remote calls (e.g., weekly video conference calls structured around specific enhancements; monthly leadership calls to discuss the overall progress of the engagement).
2) Learn about the CI challenges, successes, and CI development and management procedures within the LF. Evaluate the approach it has taken for its current and proposed CI enhancements. During the evaluation process, identify any capabilities that the LF has developed as solutions, and explore both positive aspects that can be generalized to other LFs, as well as identify aspects that need improvement.
3) Provide expertise in a number of areas, such as workflow management, networking, virtualized environments, large-scale CI deployment, data management, data analytics, gateways, and CI deployment and evaluation. This expertise can be applied to the different LF CI development areas. The Pilot can propose solutions and provide advice to the LF regarding areas of interest. As part of this effort, the Pilot can help develop and evaluate prototypes. When necessary, the Pilot can also help evaluate the LF's technical solutions in regard to cybersecurity, performance, scalability, fault tolerance, and usability.
4) Distill best practices. The engagement with the LF is bi-directional. In some cases, the Pilot can provide input to the LF; in others the LF's current practices can inform the Pilot. The Pilot can then apply this knowledge in the context of other community efforts or solutions and distill best practices that can be applicable across multiple LFs or CI projects.
5) Disseminate the knowledge gained by the Pilot during the engagement process to the broader CI community and to other appropriate technical and community forums. This knowledge can relate to both technical and social aspects of CI.
6) Foster a CI community by exploring opportunities for collaboration with other CI practitioners, projects, and community efforts.

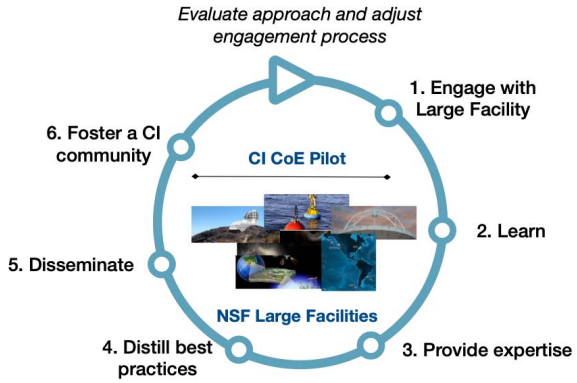

Fig. 1. Engagement with a single project.

Throughout the engagement process, we need to evaluate the effectiveness of the engagement against metrics (such as the successful development and deployment of the proposed CI solutions by the LF) and collect feedback from the LF collaborators on the usefulness of the engagement. This evaluation should be conducted periodically to ensure the effectiveness of the engagement and to monitor the goals of the engagement. Based on the evaluation, the Pilot then iterates over steps 1–6, as necessary.

IV. Working Groups: Organizing the Pilot Effort

To initiate the engagement with NEON, the Pilot held a number of virtual discussions and online teleconferences, received and reviewed a number of materials from NEON, and held a kick-off meeting at NEON's headquarters in Boulder, Colorado. During this meeting, the Pilot described the project, its goals, and in-house expertise. NEON provided an overview of their main project and identified areas in which they were interested in engaging (i.e., sensor configuration and data capturing; data collection, movement, processing, and storage; data access and dissemination; and disaster recovery). This initial meeting, which included a significant number of

CI practitioners from both projects, also established a very positive and productive line of communication between NEON and the Pilot. This positive and communicative relationship underscores the importance of in-person interactions when building productive collaborations.

The CI CoE Pilot worked with NEON to identify common challenges and standardized practices in order to aid and support NEON's specific CI needs in the most efficient and useful ways possible. An aspect of the Pilot's approach to maximizing the engagement's efficiency was the creation of topical working groups. We initially organized our Pilot project into seven working groups based on our understanding of the data life cycle within a facility and the specific needs of NEON. These specialized working groups included team members (from both projects) that could provide the knowledge and experience necessary to yield the desired results and standardize practices within a particular area. Each group was lead by two people, one from each project. The seven working groups were:

1) Data capture
2) Data storage, curation, and preservation
3) Data processing
4) Data access, visualization, and dissemination
5) Data life cycle and disaster recovery
6) Identity management
7) Engagement with Large Facilities.

Below are overviews of each working group along with the activities and accomplishments of each group.

A. Data Capture

The Data Capture working group collaborated with NEON scientists, hardware engineers, data pipeline software engineers, and web developers to evaluate various aspects of the sensor and data systems upgrade that NEON is currently implementing. Multiple themes requiring further attention emerged from this effort as the Pilot improved its understanding of NEON's practices, goals, and needs.

Technologies associated with edge and fog computing, internet of things, sensor miniaturization, and scalable data transport in less than ideal environments change at a rapid rate. This can make it difficult for individual LFs and CI projects to keep track of and evaluate advances in these areas. The CI CoE Pilot can help keep track of new technologies, prototype and evaluate latest solutions, and disseminate this information to the community.

In case of NEON, the primary focus of the Data Capture working group was the identification of appropriate and best practice technology stacks for capturing and transporting data from sensor front-ends to centralized processing and storage locations. In order to explore potential options, the working group created prototypes that demonstrated the value of selected potential technologies and tools: the use of OGCs SensorThings [18], MQTT [19] (a lightweight messaging protocol for small sensors and mobile devices that is optimized for unreliable networks), and the benefits of a full operating system and embedded systems deployment infrastructure for sensor nodes.

B. Data storage, curation, and preservation

This group, which is related to both the Data Capture and the Data access, visualization, and dissemination groups focused primarily on improving the machine readability of NEON data, which would enhance data discoverability, provenance capture, accessibility, and reusability of the data in the long term. Proactively annotating these data at the point of capture (as opposed to retroactively at later workflow stages) with community-accepted formal ontologies while adhering to community-adopted best practices reduces the chance of loss and error and improves the community's ontological quality. Because of the relationship between NEON's data and data collected by other projects in the broader community, the Data storage, curation, and preservation (DSCP) group is: 1) working with Science-on-Schema [20], a community effort focused on expanding schema.org, to appropriately accommodate scientific data, 2) collaborating with community leads in the field of CI to develop an ontological concept for a research site that is a physical entity akin to the concept of "place" in schema.org (within NEON and other projects) that hosts a number of related sensors, and 3) working with NEON staff to gather and understand the use cases that dictate which vocabulary terms and metadata need exposure for the purposes of machine readability and interoperability. Though the DSCP group's efforts were initiated as a result of discussions with NEON, it quickly became clear that such work has the potential to significantly impact other CI projects and communities (such as Earth Science Information Partners [21], EarthCube [22], Research Data Alliance [23]) as well. Thus, we are actively fostering topical discussions with these communities.

C. Data Processing

At the start of its engagement with the Pilot, NEON was leveraging and exploring the latest commercial solutions for their data processing pipelines (i.e., Airflow [24] and Pachyderm [25]). This provided the Data Processing working group with the opportunity to collect and assess NEON's knowledge and insights in the area of systematic sensor data processing and share relevant insights with other LFs and the community. In order to evaluate the applicability of the various workflow management systems to a specific scientific domain, the Data Processing group modelled existing scientific workflows in a curated selection of popular workflow management systems (WMS): Makeflow [26] and Pegasus [27], in addition to Airflow and Pachyderm. The Data Processing group is currently evaluating the results of its comparison study and aims to jointly publish an experience paper between NEON and the Pilot that compares and contrasts the different systems that were selected to model existing scientific workflows. The goal of this experience paper is to provide LFs with a reader-friendly resource guide on WMS selection by highlighting the

different strengths and capabilities of each WMS explored by the Data Processing group.

D. Data Access, Visualization, and Dissemination

The Data Access, Visualization, and Dissemination group is working on a prototype web portal, which allows for interactive exploration, easy downloading, and simple sharing of very large volumes of image data belonging to NEONs Airborne Observation Platform (AOP). The download can be requested at different resolutions and will generate an image (.png) or binary file depending on the datatype of the original data product (e.g., vegetation indices are stored as float arrays, while orthomosaic images are made of RGB data). The portal will also allow for data sharing with an auto-generated link. To support this work, the Data Access, Visualization, and Dissemination group performed data conversions (i.e., using ad-hoc scripts) of some of NEON AOPs data products (e.g., over 90 data sets of high-resolution orthorectified camera imagery mosaic, with sizes varying from 10 to 300 GB each) to a hierarchical multi-resolution data format [28]. The Data Access team then instituted a streaming server to allow for data streaming of varying data resolutions [29]. The data-streaming service and web interface have been deployed on a University of Utah server and integrated into the NEON experimental data portal. This integration required the implementation of a discovery API, which is now used by the NEON data portal to identify which data set can be explored using the interactive viewer. The discovery API provides information about site and month availability and parameters to configure the interactive viewer for the selected data set. The interactive viewer is embedded using an *iframe* providing the flexibility to use this same component in different web UI configurations (e.g., modals, windows). NEON has used this API to generate a navigation interface that allows users to select a specific site and flight (indicated by "year/month") and populate the *iframe* accordingly with the interactive viewer (see Figure 2).

More recently, the Data Access group has been working on the integration of multiple tile maps services (e.g., Google Earth) with NEON AOP data in order to deploy a visualization solution that provides a geographical context for the data collected by NEON. This requires significant efforts on the part of server data management infrastructure to fetch and combine different "tiles" from different sources into one comprehensive visualization solution. This work is particularly compelling, as the AOP data constitutes 70% of all data sets hosted by NEON by storage. Prior to these enhancements, NEON users were forced to build ad-hoc tools to visualize this AOP data. This prototypical web portal, once deployed in production, will dramatically lower the human cost of using NEON AOP data and facilitate the effortless search and retrieval of relevant datasets.

E. Data Life Cycle and Disaster Recovery

The Data Life Cycle and Disaster Recovery group has been working to:

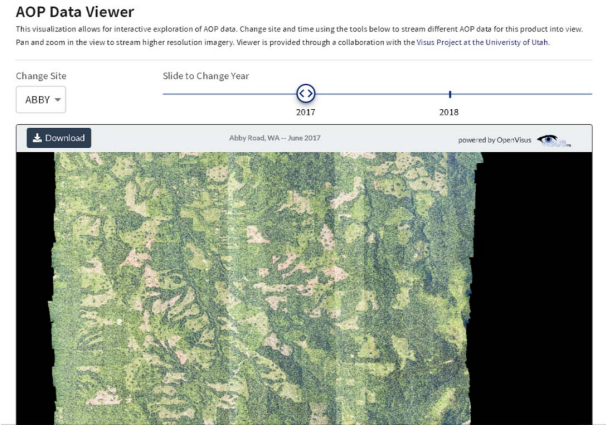

Fig. 2. AOP Interactive Web Viewer

1) understand and document the best practices and CI solutions for NEON's data life cycle (DLC) and disaster recovery (DR) methods and
2) develop effective guides and processes for DR planning across LFs. Though current versions of these guides are a direct result of the Pilot's engagement with NEON, these DR guides are in the process of being standardized for general utility and applicability (for use in other LFs and large-scale CI entities).

The LF DLC is a general model that captures the various stages that data must go through in a LF and the CI that supports the various stages of data operations. This group worked with NEON to understand the different stages of NEON's DLC. The group has documented the services and functions required for each stage of NEON's DLC and has captured the best CI practices and architectures to support each DLC stage. In doing so, the DLC and DR group has developed a generalized DLC model which can be applied to other LFs and CI projects, as depicted in Fig. 3.

Since LFs deliver data to large numbers of scientists and the public, and are entrusted to host this data for decades, disaster recovery (DR) is a cross-cutting issue across all stages of data life cycle, and effective planning for DR is essential in LF CI. Thus, the DLC and DR group identified and acted on the opportunity to start a dialogue with other LFs, such as IceCube [9], and began to develop general guidelines and effective process guides for DR [30]. These guides build upon existing federal guidelines for disaster recovery, specifically upon the National Institute of Standards & Technology (NIST) guidelines (NIST-800-34r1) [31]. Adhering to federal guidelines ensures that the Pilot project's DR template is nuanced, law-abiding, and useful. The template and planning guides will individually assist LFs in thoroughly planning for DR by performing a business-impact analysis on DR requirements and designing contingency strategies in the CI architecture for each DLC stage. The DR template and guides, once finalized, will be of use not only to LFs, in general, but also to the CI community, as a whole.

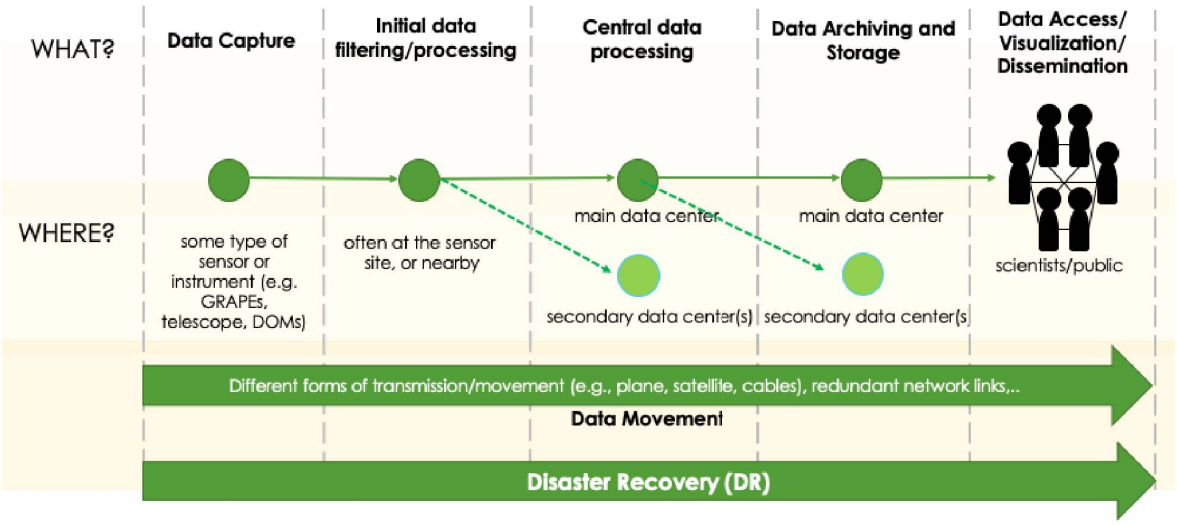

Fig. 3. Generalized Data Life Cycle of a Large Facility.

F. Identity Management

The Identity Management (IdM) group has supported NEON through the evaluation, selection, and implementation of a new identity management solution for NEON's data portal. The IdM group has taken an "advise and support" approach to keep ownership of the systems and policies within NEON, and to let NEON enhance their identity management expertise in-house. This ensures that NEON develops the resources necessary to maintain the implementation of their data portal in the long-term.

As part of their work, the IdM group produced a set of recommendations for NEON to aid in NEON's selection and implementation of identity management technologies for their data portal. Since then, the IdM group has worked closely with NEON to provide ongoing support in the adoption and integration of technologies based in OpenID Connect [32]. As a result, NEON is deploying a user-friendly authentication interface that empowers users to log in to the portal using their campus credentials via CILogon [33] or certain commercial providers (such as an ORCID). The IdM group and NEON are in the process of jointly drafting an experience paper about the data portal project with the goal of presenting the joint research at the upcoming 2019 NSF Cybersecurity Summit [3] to spread the acquired knowledge to the rest of the NSF major facilities community and to the CI community, in general.

G. Engagement with Large Facilities

The purpose of the Engagement with Large Facilities working group is to facilitate and guide interactions between LFs and the CI CoE Pilot project. A crucial outcome of this group is the organization and establishment of interaction procedures. For example, the Engagement group has developed a categorization process for different types of interactions and engagements with facilities and other large-scale CI projects to maximize benefits for both the LF and the Pilot (see Section V). The Engagement group has also developed an engagement template [34] that formalizes the engagement between the Pilot and the concerned LF. The engagement template defines the goal of the working group, specifies the time frame for the effort, identifies the activities that will be undertaken and the expected outcomes, and assesses resources to be used.

This group is also exploring several dissemination opportunities and avenues to gather and collate community feedback about the current and possible future work products of the CI CoE Pilot project. This is accomplished through attendance and interactions at: LF science-domain-specific conferences (e.g. American Geophysical Union, American Astronomical Society); venues that cater to discussions on cross-cutting CI issues for a specific set of large facilities (e.g. Open Science Grid [35], SciMMA [36] project meetings); CI and infrastructure community workshops and conferences (e.g., Practice and Experience in Advanced Research Computing (PEARC), The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)), and other NSF community events pertaining to LFs (e.g., 2019 LF Workshop [37], 2019 LF CI Workshop [38]). The Pilot project has also organized and presented webinars to disseminate best practices about cross-cutting CI issues (e.g., DLC and DR Best Practices for Large Facilities webinar [39]), and has systematically started to catalog relevant information about the CI that underpins a selected set of LFs.

In the second year of the project, the Pilot will intensify its community-building efforts and explore potential avenues to pursue. Based on a review of existing research-based approaches to community building, the Pilot project is currently evaluating potential strategies that may be effective in the CI domain. One such strategy is the creation of a community-curated portal for the sharing of CI-related challenges and solutions. A community CI portal would provide a centralized

resource for various LF CI practitioners, and would thus offer one avenue of building and growing the CI-centered community. A potential limitation of such a centralized approach is ensuring practitioner buy-in and participation, as such a resource would only be an effective community-building tool if it is actively engaged with. Fostering the emergence of a more decentralized, network-based community around CI for LF is another potential approach. Identifying a few LFs that already exchange solutions and best practices, and thus can serve as the initial seed for such a network, would be the first step in the process, and the Pilot project would further foster such network-based community by serving as a clearing house and by connecting LFs to this growing network. Combining both approaches may be most beneficial, were we develop a Pilot-based CI catalog and help build a network of such resources.

V. Engagement Methodology

During its engagement with NEON, the Pilot also defined other types of potential interactions with the LFs and other large-scale CI projects. Below, we briefly describe the three main types of engagement: 1) deep engagements (the work with NEON being an example), 2) topical discussions, and 3) community building. The definitions and characteristics of each type of engagement are explained in greater detail below.

Deep engagement interactions occur when the Pilot can work with a particular facility to identify an important topic or topics that require(s) joint problem-solving. For each topic the LF and Pilot form a topical working group composed of members of each project. The group leads are also identified (one from each project team). The deep engagement strives to conduct focused discussions to better understand the identified challenges, to leverage in-house and community CI knowledge, and to employ and evaluate efficient solutions. Deep engagement combines virtual and in-person meetings to maintain an effective and reliable stream of communication between the facility and the CI CoE Pilot. Deep engagements need to establish consensus on the general timeline of the collaboration and they need to define specific products. A deep engagement can be defined to last a month or several months depending on the complexity of the tasks, the timeline of the two projects, and the availability of resources. Deep engagement interactions can yield outcomes such as documents and papers, presentations and videos of webinars, templates and guides, prototypes, schema implementations, and/or demonstrations.

Topical discussion interactions occur when the Pilot is able to identify a topic that is of significant importance to multiple facilities. Topical discussion interactions entail the facilitation of virtual discussions across a number of large facilities. This involves presentations and discussions on the identified topic during conference calls and webinars, at topical workshops, conferences, and community-based events; the collection and sharing of experiences and best practices; and distilling and establishing best practices and lessons learned for the identified CI topic. The outcomes of a topical discussions can consist of standardized templates and guides that can be widely applied to various LFs. Some products (such as schemas) can also be contributed to other community efforts (for example, to the Earth Science Information Partners (ESIP) in the case of the schema.org effort– Section IV-B). Topical discussions can also lead to closer collaboration between LFs, for example in the area of shared services. The outcome of a topical engagement is increased collaboration and problem-solving across facilities through greater awareness of CI experiences, practices, and solutions. Although topical discussions are also defined to last a specific amount of time, they are meant to last over a period of months and their timelines are not as strict as that of the deep engagements.

Community building efforts aim to build a community around cyberinfrastructure. Today there are a number of disconnected interest groups within the CI landscape and it can be hard for CI practitioners to connect to their peers across projects and groups. Thus, the Pilot aims to help connect existing groups into a broader CI network. The Pilot aims to bring in new members to the CI CoE Pilot effort and to reach out to other communities to enable sharing experiences and knowledge. Community building engagements are meant to last throughout the lifetime of the project.

Community building efforts also include collecting and disseminating information about the broad CI community activities, both technical and social, such as workforce enhancement and retention. We recognize that significant effort in the area of community building across the CI workforce is done within the research computing centers on campuses; thus, we have initiated discussions with the Campus Research Computing Consortium (CaRCC) [40] to explore potential areas of collaboration.

As the first year of the project comes to a close, the CI CoE Pilot is evaluating its engagement and experiences with NEON and determining next steps. During its second year, the Pilot plans to engage with additional LFs, appraise the best model for engagement with LFs, distil best CI practices, and develop training and prototypical demonstrations using advanced CI technologies. We also plan to continue to identify related efforts and build a community around CI.

VI. Partnerships

The Pilot has partnered with Trusted CI (formerly CTSC) [4], which has been an important, independent resource for LFs and large cyberinfrastructure projects in the area of cybersecurity. Trusted CI has shared its experience in and process for engagement planning, as well as practices for building connections within LFs, with the Pilot. This greatly reduced the time required by the Pilot to spin up functional engagement programs and allowed the Pilot to start producing results for NEON more quickly.

Just as TrustedCI provides leadership in the cybersecurity arena, the Pilot aims to provide leadership in the area of robust, production-quality cyberinfrastructure, and we are learning about TrustedCI's practices and engagement processes. For example, our engagement template is closely modeled on

the one developed by Trusted CI. In order to support this collaboration with Trusted CI, we are co-funding personnel between the two projects.

We have also developed partnerships with the Open Science Grid [35], a large-scale, high-throughput computing community platform, and the Science Gateways Community Institute [41], an NSF Software Institute. Members of these large CI projects are part of the advisory board of the Pilot effort and are providing us with advice based on the experience they have gained over the years of serving their communities. Additional advisory committee members include representatives of LFs and large CI projects, as well as CI experts [5].

VII. Conclusions and Lessons Learned

Although we have learned a number of technical lessons, from understanding NEON's CI architecture and infrastructure to discovering new workflow management tools and capturing the end-to-end data life cycle, many of the lessons learned were in the area of project organization/management and the importance of social aspects of collaborations.

On the project management side, the adoption of the concept of working groups that focus on particular topics helped organize our teams and enabled us to define manageable goals and keep track of progress over time. Having a well-defined engagement plan for each working group (based on the template) was also important, as it set expectations for the interactions and formalized the expected outcomes for each team. Based on this understanding, during an in-person meeting between the Pilot and NEON in August of 2019, we were able to sum up the various working group products and declare completion on five out of the seven working groups. We also decided to re-activate the Identity Management working group to work on managing security tokens for APIs used to access NEON data.

The success of the engagement with NEON also depended on good timing. As the Pilot was starting out, NEON was entering its enhancement phase, which made NEON receptive to collaborating with the Pilot on the technical CI challenges they were facing. The overlapping of NEON's existing enhancement timeline with the Pilot's engagement timeline fostered the rapid pace of the engagement and the rich flow of ideas and information between the two projects. In some areas, such as data collection and processing, NEON already had significant experience and expertise and was able to share this knowledge with the Pilot. In turn, the Pilot was able to synthesize the information, augment it with its own experiences, and disseminate the results (as in the case of the WMS comparison study). In other areas such as identity management and visualization, the Pilot's expertise directly contributed to NEON's enhancements goals, adding resources to its effort.

Our collaboration with NEON also illuminated the need to form personal relationships between the projects' participants. Although we had productive conference calls between the two projects, better outcomes and more in-depth discussions were enabled by in-person meetings. Based on interactions during such meetings, breaks, and social events, the Pilot and NEON started building a rapport and sense of trust, which also translated to more effective virtual interactions. We believe that other successful engagements with LFs will also require this important inter-personal effort.

Since NEON was the first target of engagement for the Pilot, we will refine our engagement strategies to scale the approach to other LFs and the broader CI community. We have already started engaging other LFs (such as IceCube [9] and OOI [15]) in the area of the data life cycle to understand whether the Pilot's model is sufficient to represent the data life cycle of other LFs. The next steps will be to map this life cycle to the CI services that support it within various LFs and to conduct an analysis of the solutions used and potential areas of collaboration and CI re-use.

Ultimately, the goal of the Pilot effort is to develop a model and a blueprint for a CI CoE that will serve as a platform for knowledge sharing and community building around CI for LFs and other large-scale CI projects We hope that such a CI CoE will become a key partner for the establishment and improvement of LFs with advanced CI architecture designs and provide a trusted forum for discussions about CI sustainability and workforce development, training, and retention.

Acknowledgments

This work was supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer and Information Science and Engineering and the Division of Emerging Frontiers in the Directorate for Biological Sciences Under Grant #1842042. The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the National Science Foundation through the NEON Program. We acknowledge the discussions with and contribution of the Earth Science Information Partners (ESIP) and the members of the ESIP Semantic Technologies Committee. The Pilot would also like to thank is Advisory Committee for their guidance.

References

[1] M. Parashar, S. Anderson, E. Deelman, V. Pascucci, D. Petravick, and E. M. Rathje, "2017 NSF Large Facilities Cyberinfrastructure Workshop," 2017. [Online]. Available: http://facilitiesci.org/assets/reports/facilitiesci-workshop-report-11-17.pdf

[2] S. Russelland, C. Jackson, B. Cowles, and K. Avila, "2017 NSF Community Cybersecurity Benchmarking Survey Trusted CI Report ," 2017. [Online]. Available: http://hdl.handle.net/2022/22171

[3] "2019 Trusted CI Cybersecurity Summit ," 2019. [Online]. Available: https://trustedci.org/2019-nsf-cybersecurity-summit

[4] A. Adams, K. Avila, J. Basney, D. Brunson, R. Cowles, J. Dopheide, T. Fleury, E. Heymann, F. Hudson, C. Jackson, R. Kiser, M. Krenz, J. Marsteller, B. P. Miller, S. Piesert, S. Russell, S. Sons, V. Welch, and J. Zage, "Trusted CI Experiences in Cybersecurity and Service to Open Science. PEARC'19: Practice and Experience in Advanced Research Computing," Tech. Rep., 2019. [Online]. Available: https://doi.org/10.1145/3332186.3340601

[5] (2019) Cyberinfrastructure Center of Excellence Pilot. [Online]. Available: https://https://cicoe-pilot.org/

[6] National Science Foundation, "NSF Research Infrastructure Projects ," 2019. [Online]. Available: https://www.nsf.gov/bfa/lfo/docs/major-facilities-list.pdf

[7] R. Perley, C. Chandler, B. Butler, and J. Wrobel, "The Expanded Very Large Array: A new telescope for new science," *The Astrophysical Journal Letters*, vol. 739, no. 1, p. L1, 2011.

[8] Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew *et al.*, "LSST: from science drivers to reference design and anticipated data products," *The Astrophysical Journal*, vol. 873, no. 2, p. 111, 2019.

[9] M. Aartsen, M. Ackermann, J. Adams *et al.*, "The IceCube Neutrino Observatory: instrumentation and online systems, journal = Journal of Instrumentation," vol. 12, no. 03, pp. P03 012–P03 012, mar 2017. [Online]. Available: https://doi.org/10.1088%2F1748-0221%2F12%2F03%2Fp03012

[10] C. Collaboration *et al.*, "The CMS experiment at the CERN LHC," 2008.

[11] G. Aad, J. Butterworth, J. Thion, U. Bratzler, P. Ratoff, R. Nickerson, J. Seixas, I. Grabowska-Bold, F. Meisel, S. Lokwitz *et al.*, "The ATLAS experiment at the CERN large hadron collider," *Jinst*, vol. 3, p. S08003, 2008.

[12] O. S. Board, N. R. Council *et al.*, *Science at sea: meeting future oceanographic goals with a Robust Academic Research Fleet*. National Academies Press, 2009.

[13] G. M. Harry, L. S. Collaboration *et al.*, "Advanced LIGO: the next generation of gravitational wave detectors," *Classical and Quantum Gravity*, vol. 27, no. 8, p. 084006, 2010.

[14] I. Rodero Castro and M. Parashar, "Architecting the cyberinfrastructure for National Science Foundation Ocean Observatories Initiative (OOI)," in *Instrumentation viewpoint*, no. 19. SARTI, 2016, pp. 99–101.

[15] I. Rodero and M. Parashar, "Data Cyber-Infrastructure for End-to-end Science: Experiences from the NSF Ocean Observatories Initiative," *Computing in Science Engineering*, pp. 1–1, 2019.

[16] "NEON National Ecological Observatory Network ," 2018. [Online]. Available: https://www.neonscience.org/observatory/about

[17] D. T. Barnett, P. A. Duffy, D. S. Schimel, R. E. Krauss, K. M. Irvine, F. W. Davis, J. E. Gross, E. I. Azuaje, A. S. Thorpe, D. Gudex-Cross *et al.*, "The terrestrial organism and biogeochemistry spatial sampling design for the National Ecological Observatory Network," *Ecosphere*, vol. 10, no. 2, p. e02540, 2019.

[18] S. Liang, C.-Y. Huang, and T. Khalafbeigi, "OGC SensorThings API Part 1: Sensing, Version 1.0." 2016.

[19] U. Hunkeler, H. L. Truong, and A. Stanford-Clark, "Mqtt-sa publish/subscribe protocol for wireless sensor networks," in *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*. IEEE, 2008, pp. 791–798.

[20] E. S. I. Partners, "Provides guidance for publishing schema.org as JSON-LD for the sciences ," 2019. [Online]. Available: https://github.com/ESIPFed/science-on-schema.org

[21] "Earth Science Information Partners," 2019. [Online]. Available: https://www.esipfed.org/

[22] "Earthcube," 2019.

[23] "Research Data Alliance," 2019. [Online]. Available: https://www.rd-alliance.org/

[24] M. Beauchemin. (2014) Apache Airflow Project. [Online]. Available: https://airflow.incubator.apache.org/

[25] Pachyderm, Inc. (2017) Pachyderm. [Online]. Available: https://www.pachyderm.io/

[26] M. Albrecht, P. Donnelly, P. Bui, and D. Thain, "Makeflow: A portable abstraction for data intensive computing on clusters, clouds, and grids," in *Proceedings of the 1st ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*. ACM, 2012, p. 1.

[27] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, "Pegasus: a Workflow Management System for Science Automation," *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.

[28] S. Kumar, V. Vishwanath, P. Carns, B. Summa, G. Scorzelli, V. Pascucci, R. Ross, J. Chen, H. Kolla, and R. Grout, "PIDX: Efficient parallel I/O for multi-resolution multi-dimensional scientific datasets," in *IEEE International Conference on Cluster Computing*, 2011.

[29] V. Pascucci, G. Scorzelli, B. Summa, P.-T. Bremer, A. Gyulassy, C. Christensen, S. Philip, and S. Kumar, "The ViSUS Visualization Framework," in *High Performance Visualization: Enabling Extreme-Scale Scientific Insight*, E. W. Bethel, H. Childs, and C. Hansen, Eds. CRC Press, 2012.

[30] (2019) Cyberinfrastructure Center of Excellence Pilot Templates. [Online]. Available: https://cicoe-pilot.org/materials/templates

[31] National Institute of Standards and Technology (NIST), "Contingency Planning Guide for Federal Information Systems ," 2010. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-34r1.pdf

[32] "the web sso standard openid connect: In-depth formal security analysis and security guidelines."

[33] J. Basney, T. Fleury, and J. Gaynor, "CILogon: A federated X. 509 certification authority for cyberinfrastructure logon," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 13, pp. 2225–2239, 2014.

[34] Cyberinfrastructure Center of Excellence Pilot (CI CoE Pilot), "Cyberinfrastructure Center of Excellence Pilot Engagement Plan ," 2019. [Online]. Available: https://github.com/cicoe/engagement-templates/blob/master/CICoE-Pilot-Engagement-Plan-Template.pdf?raw=true

[35] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein *et al.*, "The Open Science Grid," in *Journal of Physics: Conference Series*, vol. 78, no. 1. IOP Publishing, 2007, p. 012057.

[36] (2019) Scalable Cyberinfrastructure to support Multi-Messenger Astrophysics. [Online]. Available: https://scimma.org/

[37] "Large facilities workshop," 2019.

[38] "2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure ," 2019. [Online]. Available: https://facilitiesci.org/

[39] (2019) Cyberinfrastructure Center of Excellence Pilot Videos. [Online]. Available: https://cicoe-pilot.org/materials/videos

[40] "Campus Research Computing Consortium (CaRCC)," 2019. [Online]. Available: https://carcc.org

[41] N. Wilkins-Diehr and T. D. Crawford, "NSFs inaugural software institutes: The science gateways community institute and the molecular sciences software institute," *Computing in Science & Engineering*, vol. 20, no. 5, pp. 26–38, 2018.