

“THEME ARTICLE”, “FEATURE ARTICLE”, or “COLUMN” goes here: The theme topic or column/department name goes after the colon.

# The Virtual Data Collaboratory

## A Regional Cyberinfrastructure for Collaborative Data-Driven Research

**Manish Parashar**  
Rutgers University

**Vasant Honavar**  
Pennsylvania State University

**Anthony Simonet**  
Rutgers University

**Ivan Rodero**  
Rutgers University

**Forough Ghahramani**  
Rutgers University

**Grace Agnew**  
Rutgers University

**Ron Jantz**  
Rutgers University

The Virtual Data Collaboratory is a federated data cyberinfrastructure designed to drive data-intensive, interdisciplinary and collaborative research that will impact researchers, educators and entrepreneurs across a broad range of disciplines and domains as well as institutional and geographic boundaries.

Scientific progress across disciplines is increasingly enabled by our ability to examine natural phenomena through the computational and/or data-centric lens (e.g., using algorithmic or information processing abstractions of the underlying processes) and our ability to acquire, share,

integrate, steward, and analyze disparate types of data [1]. Multi-million-dollar projects and instruments such as the Large Synoptic Survey Telescope (LSST) in Chile, the Large Hadron Collider (LHC) in Switzerland, the Ocean Observatory Initiative (OOI) [2] in the United States are all producing, or will produce, Peta Bytes of open-use data every year. Storing, filtering, analyzing and more generally transforming these large datasets into knowledge require large high-end cyberinfrastructures capable of keeping up with the pace of streaming data. At the same time individual researchers and smaller research groups are also generating raw data and data products, which also need to be managed, analyzed and shared.

However, the data and infrastructure necessary to support this data-driven transformation of science is largely missing: while a few projects (such as those mentioned above) can afford a dedicated cyberinfrastructure, these are limited to distributing raw data and data products from the projects, but do not offer support to other smaller research groups. Within the next decade, a global, integrated data science infrastructure will be essential for scientific and scholarly discov-

ery. Designed properly, this infrastructure will free researchers from trying to manage, manipulate, process, share, and preserve large datasets (often residing in siloed environments), and allow them to concentrate on research by simplifying the process of extracting scientific meaning from theoretical, experimental, and observational data. Not only it is critical that the infrastructure ensures the reproducibility of data-driven science, but that it is highly configurable, extensible, and sustainable.

This article introduces The Virtual Data Collaboratory (VDC), a project funded by the National Science Foundation (NSF) that aims at building such a cyberinfrastructure across Rutgers University (RU) in New Jersey and Penn State University (PSU) in Pennsylvania, with the potential to incorporate additional research institutions across the nation. The overarching goal of VDC is to transform shared data as a core modality for research and discovery. VDC is a federated data cyberinfrastructure that is designed to drive data-intensive, interdisciplinary and collaborative research and enable data-driven science and engineering discoveries. VDC accomplishes this goal by providing seamless access to data and tools that are part of the federation to researchers, educators, and entrepreneurs across a broad range of disciplines and scientific domains as well as institutional and geographic boundaries. VDC will lead the way to collaborative and open data science by providing methods, software and hardware tools to be reused by cyberinfrastructures worldwide.

The overarching goal of VDC is to transform shared data as a core modality for research and discovery.

Central to the VDC vision are three infrastructural innovations: (i) regional science DMZ (Data DMZ) that provides the data import/export services and necessary services to enable efficient and transparent access to data and computing capabilities regardless of scientist location, (ii) expandable and scalable architecture for data-centric infrastructure federation that supports peer-to-peer federation while respecting local constraints, and (iii) a data services layer collaboratively built to support research workflows that utilize cutting-edge semantic web technologies and an innovative collaboration and connection service layer to support interdisciplinary research, expand access, and increase the impact of data-science worldwide. The data services layer fully integrates the researchers and tools that create data with the data itself, as part of the discovery and reuse process. To ensure the durability and citability of deposited data, the data services will attach a Digital Object Identifiers (DOIs) to each object. The data services will manage the long-term data lifecycle, ensure the immutability and authenticity of data, in order to create reproducible research today and years ahead.

VDC brings together a deeply engaged interdisciplinary team of researchers and infrastructure organizations to build a next-generation data-centric cyberinfrastructure that promotes collaboration and identifies relationships among research products to facilitate deep and intuitive reuse of research data. VDC is led by the Rutgers Discovery Informatics Institute (RDI2) (a university-wide institute focused on compute- and data-intensive research across science and engineering), and other Computer Science and Library groups from Rutgers University and PennState University. In partnership with KINBER and NJEdge, respectively the Pennsylvania and New Jersey Research and Education Networks, and the New Jersey Big Data Association (NJBDA), an alliance that unites universities and colleges across New Jersey, we will explore expanding this collaboratory to additional institutions in New Jersey, Pennsylvania, and beyond.

VDC will also outreach to education bodies in order to train the next generation of scientists with deep interdisciplinary expertise, a high degree of competence in large-scale data science and collaborative and open science.

The project has been organized into six key focus areas, each involving researchers from the participating institutions, including Rutgers and PSU:

- Networking, focused on designing, deploying and testing the Data DMZ and associated services.

- Systems, focused on data and analytics infrastructure and software stack at the two hubs, as well as the federation for satellite sites and remote data stores.
- Data Services, focused on developing an interface which supports interdisciplinary discovery of the data most useful to a researcher's need.
- Application Integration, focused on integrating the use cases (VDC-SBUC and VDC-OOI) into the VDC computing and federation infrastructure, informing the design, implementation, and evaluation of the VDC system.
- Outreach and broader impacts, focused on improving the training of scientists, engineers, and entrepreneurs on large federated data systems like VDC and bringing diverse participation to the learning and practice of data analytics using VDC.

This article offers an overview of the approach VDC is pursuing to construct the next generation of integrated cyberinfrastructures for data science, and of the status of work in progress in the above focus areas.

## SCIENTIFIC CYBERINFRASTRUCTURES

VDC leverages technologies and software components developed across a range of fields including networking, data analytics, data cataloging and provenance. In addition, the VDC networking infrastructure and data transfer services are based on Flash I/O Appliances (FIONAs). FIONA is a hardware and software specification proposed by the University of California at San Diego for building affordable Data Transfer Nodes (DTNs) out of commodity servers.

Multiple projects in the past have contributed to making research more collaborative through domain-specific approaches. For example, the DataOne [3] project focuses on earth sciences and the NEON [4] project focuses on ecological research. Recent projects funded by the NSF under its Data-Infrastructure Building Blocks (DIBBs) program offer approaches that are more generic and can contribute to multiple scientific domains. The DataCenterHub project [5] develops a web platform for uploading, sharing and discovering datasets; the platform can record user-defined properties and metadata as well as provenance information, linking experiment objects to datasets and files, but does not provide computing resources or analytics software. CyberGIS [6] focuses on spatial data and aims to offer a cloud-based platform wherein scientists can publish both data and operations that apply to the data in a way that favors reusability and stable performance over time. The SeedMe project [7] provides a collaborative space for researchers to exchange transient data and preliminary results with integrated visualization tools and Application Programming Interfaces (APIs) for interfacing with High Performance Computing jobs. Its features are thus similar to the embargo feature of VDC that allows data to be kept private and shared amongst a select group of users, but it lacks close integration with compute resources that VDC aims to provide.

Some projects that provide computing infrastructures for science could join the VDC federation, benefitting users from all collaborating projects. This is the case for the Pacific Research Platform (PRP) project [8] that integrates regional campus science DMZs and computing resources including GPU nodes into a high-capacity data-sharing infrastructure. The Aristotle Cloud Federation [9] aggregates resources from NSF resources such as Jetstream as well as public cloud resources such as the Amazon Web Services in order to alleviate the effort required to select and use cloud resources. VDC will be the first generic and interdisciplinary cyberinfrastructure providing both a collaborative environment for discovering, cataloging and sharing data as well as networking, computing resources and software for any scientific domain.

## VDC VISION AND OBJECTIVES

VDC aims to conceptualize, design and provide a blueprint for future state of the art data-intensive cyberinfrastructure that is based on the federation of computing, storage and networking equipment and an innovative data services layer. It will be a prototype infrastructure federated across three geographically distributed Rutgers University (RU) campuses in New Jersey (Camden, Newark and New Brunswick), multiple campuses in Pennsylvania (Pennsylvania State University (PSU), Drexel University, Temple University and the University of Pittsburgh) and

beyond (City University of New York (CUNY)). All campuses will be coupled by a high-speed network managed by two Regional Education and Research Networks, New Jersey's (NJEdge) and Pennsylvania's KINBER, with the potential to expand by incorporating academic/research institutions across the Mid-Atlantic and the nation. The VDC will build on and integrate with existing regional, national, and international data repositories (including NSF funded repositories like the Ocean Observatories Initiative (OOI) and the Protein Data Bank (PDB)), and leverage local/regional/national ACI investments, such as the NSF funded Pacific Research Platform (PRP), Big Data Regional Hubs, XSEDE, the Open Science Grid (OSG) [10] and Campus Cyberinfrastructure projects. This infrastructure will allow the integration of these current (and future) frameworks through a virtual collaboratory that can be accessed by researchers, educators, and entrepreneurs across institutional and geographic boundaries, stimulating community engagement and accelerating interdisciplinary research. Additionally, we will develop online learning modules to support STEM education initiatives.

When completed, the prototype infrastructure will provide data scientists and researchers with services and features that will: connect the participating campuses with a high-performance network for faster data exchange; allow for the integration of current and future frameworks; federate computing resources to offer researchers more computing power than their campus alone can offer; provide researchers with easy access to cutting edge big data software; make big data science more open and collaborative with a set of tools for sharing and referencing datasets; and promote big data science to students and companies through outreach programs.

## Research Challenges

Providing efficient and transparent access to data and computing capabilities regardless of scientist location leads to a number of research challenges in networking, scheduling, infrastructures, data management and provenance.

A first challenge is expandability and scalability: the VDC must accommodate new partner institutions without impacting other partners. Further, the geographic distribution of resources must increase the performance of the whole infrastructure —by exploiting data locality and reducing network traffic— and the pool of resources available to all researchers. Efforts will focus on making the federation function in a peer-to-peer fashion, while respecting the local constraints of each partner. The heterogeneity of the resources that compose the federation comes as an additional challenge; external data repositories relying on different software stacks need to be searchable and connected to the federation; computing and storage resources on all partner campuses need to be accessible transparently to allow for the execution of complex distributed workflows. Finally, this large heterogeneous infrastructure will be used by researchers and instructors from many fields and as infrastructure designers, we cannot expect all of them to be computer science proficient. The federation must thus be entirely transparent to VDC users.

## OVERALL ARCHITECTURE OF VDC

Figure 1 provides a schematic overview of the VDC architecture. Its foundation is the network and data infrastructure layer. The network infrastructure (Data DMZ) will implement high-bandwidth connectivity between VDC Hubs and spokes located in New Jersey and will reach the broader community via regional and national network connections, including Internet2. The data infrastructure is composed of a federated object store that integrates existing publicly accessible data stores (e.g., PDB and OOI) with newly added data. The Data Services Layer provides the software infrastructure required to support easy cataloging, connecting, persisting, and querying of the data, and research workflows and collaboration. Its three layers are briefly described here and further detailed in the System Design section.

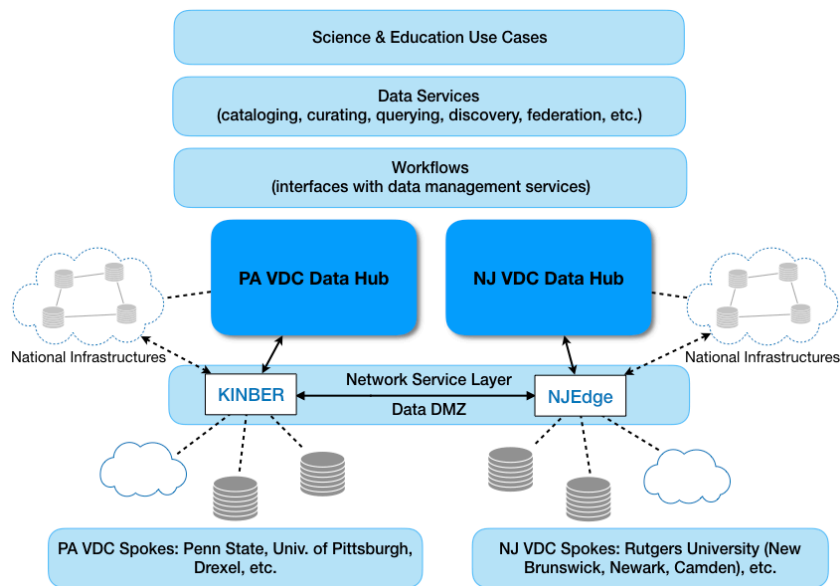


Figure 1. High-level VDC architecture

1. **The Network and Data Infrastructure layer** aims to alleviate the difficulty commonly encountered by researchers trying to manage and analyze large-scale scientific data with infrastructures not designed for the types of problems they want to solve. As shown in Figure 1, the VDC will be a federated and coordinated data solution spanning across multiple campuses within New Jersey and Pennsylvania and enabling direct access to either campus and other national resources like the Open Science Grid. The federated data infrastructure leverages geographic distribution, reducing network traffic by storing data close to where it is likely to be used. The two VDC hubs will offer a mix of storage based on the Hadoop File System (HDFS) for data analytics and of a Network-Attached Storage (NAS) staging area for file-based data ingestion and delivery. The storage solution will serve as: (i) primary storage for new data; (ii) a collection and delivery point for data already maintained elsewhere; and (iii) long-term storage for data to be archived.
2. **The Data Service Layer (DSL)** is a data management and services system that utilizes logical relationships between data elements and objects to connect researchers to information to increase research quality and interdisciplinary breadth. The DSL builds on robust, mature open-source technologies to create a repository of curated research data that is authentic, citable in the scholarly record, discoverable, and reusable. Archived data made understandable by context-based metadata is the first step of a useful repository.
3. **The Workflows layer** provides interfaces for querying the data management services. They allow the user to search for objects, metadata and relations between data objects, and provide the necessary interfaces for workflows and tools to use these objects.

## DRIVING USE-CASES

The value of VDC stems from significant impact on the science created by researchers who use it. Leading researchers at Rutgers, Penn State, and beyond have recognized the need for the VDC and have maintained a strong interest in interdisciplinary research. VDC members from Penn State and Rutgers defined early on in the project two use cases based on their own research domains—Oceanography and Bioinformatics—with two goals in mind; first, these use-cases are intended to drive the design of the VDC architecture and implementation; second, they will be used for the evaluation of the infrastructure, both in terms of features (the VDC must provide the tools necessary to run these applications) and performance (the VDC must satisfy the applications while optimizing some metrics, e.g. execution time, resource usage, network contention).

## The Ocean Observatory Initiative

Large scientific facilities provide researchers with instrumentations, data, and data products that can accelerate scientific discovery. However, while these facilities provide reliable and pervasive access to the data and data products, users typically have to download the data of interest and then process them, typically using local resources. Consequently, transforming these data and data products into insights requires local access to powerful computing, storage, and networking resources. These requirements can significantly limit the impact of the data, especially for researchers, educators, and students who do not have access to such capabilities. We are currently experiencing this limitation in the case of the Ocean Observatories Initiative (OOI) [11]. OOI currently serves data from 57 stable platforms and 31 mobile assets, carrying 1,227 instruments (~850 deployed), providing over 25,000 science datasets and over 100,000 scientific and engineering data products. OOI raw data and data products, such as high-definition video and hydrophone data, are rapidly growing in size and even modest queries can result in significant latencies for end users and can overwhelm their local storage and computing capabilities. Furthermore, researchers are exploring mechanism for combining OOI data with data from other observing system as part of their workflows.

The VDC-OOI Use Case aims to enable such data-driven end-to-end workflows, which combine data from OOI and other facilities and can leverage computing and networking resource that are part of the national cyberinfrastructure. VDC is used for automated data sharing and data processing and delivery based on data subscription and data-driven workflows. Specifically, this use case will enable users to create data streams based on queries across multiple data stores, subscribe to these streams, and associate workflows with stream and stream-related events that when triggered can seamlessly orchestrate the entire data-to-discovery pipeline. Such a pipeline will involve executing the queries on the OOI (and other) cyberinfrastructure, staging the data to VDC computing/analytics resources, launching the modeling and analysis processes to transform such data into insights, and publishing results to the user. This solution leverages the Apache Kafka data streaming platform on top of FIONA-based nodes, which can build a content delivery network (CDN) using distributed data hubs.

For this use-case we are considering applications that require data streaming from OOI and UNAVCO high-precision GPS stations to be analyzed in real time. The joint analysis of both data sources is expected to improve the delay for detecting tsunamis. In particular, it demonstrates an important feature of VDC, which is how to program and execute analysis tasks involving multiple datasets that cannot – even temporarily – be stored in a single place.

## Structural Bioinformatics

Deciphering Sequence and Structural Correlates of Protein Nucleic Acid Interactions: this VDC use case (referred to as the VDC Structural Bioinformatics Use Case (VDC-SBUC)) aims to exercise, demonstrate, and guide the further development of some of the key elements of the VDC infrastructure. In this context, the VDC will be used to create a collaborative assembly, integration, and analysis platform for several datasets of protein-nucleic acid complexes derived from the Protein Data Bank (PDB) —the archival data resource for all experimentally derived biological macromolecules and their complexes [12], the Nucleic Acid Database (NDB) —a specialized data resource containing information about nucleic acid containing structures, and related sources. Specifically, this use case aims to establish, use, and evaluate a shared data and computational infrastructure, complete with computational workflows for documenting, comparing, and reproducing computational analyses and prediction of protein nucleic acid complexes and interfaces. Examples of such analyses include characterization of conformational changes in proteins upon binding to DNA, computational prediction of protein-DNA and protein-RNA complexes, etc. The resulting datasets will be curated, assigned DOIs, versioned, indexed, and shared to support intentional revisions to data and analysis tools. The digital artifacts resulting from this use case will be linked to the work products (data, workflows) used to create them using the VDC's Data Services Layer within the shared data and computing environment of the VDC. This use case leverages the linked data capabilities of the Samvera platform to enable data to be linked to (i) the tools that created them, (ii) the intermediate data products produced by the tools (analyses,



visualizations) and (iii) the articles produced through citations referencing the data and data product DOIs.

## VDC SYSTEM DESIGN

### Storage and Analytics Infrastructure

The VDC federated storage architecture leverages geographic distribution, reducing network traffic by storing data close to where it is likely to be used. The proposed HDFS-based storage system provides reliable multi-petabyte usable storage across the two VDC hubs using data replication. HDFS provides superior performance and scalability compared to Storage Area Network (SAN)-based solutions and exploits data locality by supporting execution of data analytics on the same hardware. It also enables high reliability without requiring expensive offsite backup for fault tolerance. In addition to the HDFS-based data store, each site has a NAS-based staging area for file-based data ingestion and delivery methods. The storage solution serves as: (i) primary storage for new data; (ii) a collection and delivery point for data already maintained elsewhere; and (iii) long-term storage for data to be archived. The system presents a Network File System (NFS) interface as a conventional file interface through the NAS, as well as REST and other web services for future usage support. This HDFS-based store system scales better than NFS appliances at less expense and is particularly appropriate for the dual requirements of archiving and distributed data access and processing. In addition to offering the expandability and scalability properties that VDC requires, HDFS offers flat view of geographically distributed data and erases hardware heterogeneity, making the storage service completely transparent to the users. Finally, its programming interfaces makes it easy to integrate the storage service with existing software and to develop a set of dedicated tools for VDC specific to the VDC.

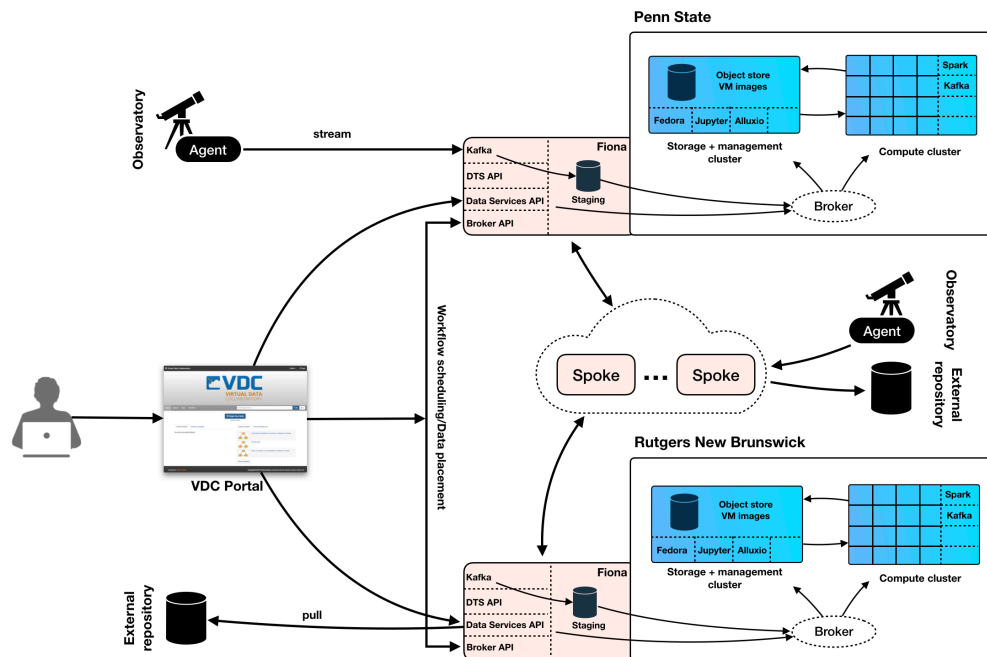
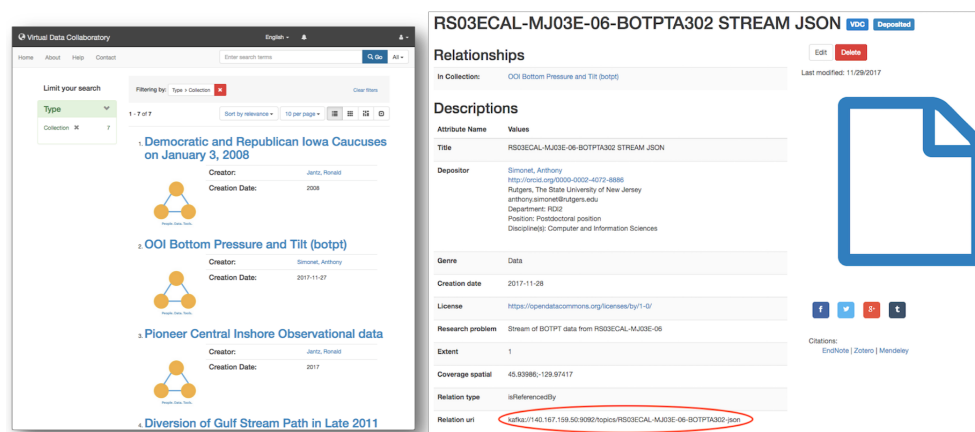


Figure 2. VDC computing and storage architecture

## Software Architecture

The architecture presented in Figure 2 shows the logical organization of hubs and spokes inside the VDC, the links between their computing and storage resources, and the main services they provide to support the use-cases presented above. Each participating institution is connected to the VDC network through a FIONA-based [13] node that hosts interconnection and user-facing services. Each hub will provide three sets of APIs that are currently under development; the Data Transfer Service (DTS) API will be an endpoint for bulk data transfers between sites; the Data Services API will expose the datasets stored locally and associated metadata; the Broker API will allow the submission, scheduling and placement of data processing jobs and resource monitoring. Users will use either of these APIs directly to upload/download datasets and to submit jobs or use the VDC portal shown in **Error! Reference source not found.** which will offer a user-friendly interface to the main services.

Additionally, FIONA-based nodes host a Kafka server and additional agents for supporting the ingestion of real-time data streaming from observatories, sensor networks, etc. These streams are staged temporarily on the FIONA-based node storage; from there, they can be streamed back to users or be processed by the compute cluster or other compute resources managed by VDC. To support the discovery of data streams, the Data Services and the VDC portal allow data records to contain either a file or a URL to a stream, as illustrated by **Error! Reference source not found.** For stream entries, the protocol part of the URL indicates which streaming software provides the resource, the host and port parts indicate where the resource is located, and the path part of the URL uniquely identifies the stream.



The figure consists of two side-by-side screenshots of the Virtual Data Collaboratory (VDC) portal. The left screenshot shows the homepage with a search bar, filters, and a list of datasets. The right screenshot shows a detailed view of a dataset titled 'RS03ECAL-MJ03E-06-BOTPTA302 STREAM JSON'. The 'Relation uri' field is circled in red, highlighting a Kafka URL.

**Figure 3. Virtual Data Collaboratory portal homepage (left) and detailed data record view (right). The circled text highlights a URL to a Kafka stream of data from the Ocean Observatory Initiative.**

At each hub the backend hosts two computer clusters. The storage and management cluster offer disk and in-memory storage with HDFS coupled with Alluxio and scientific toolkits for data exploration and collaboration such as Jupyter Notebook. The compute cluster offers more servers (the actual number varies for each hub) for performing heavier computations. There, different software stacks can be installed to accommodate various needs; typically, Apache Kafka, Spark and Hadoop are used to execute stream- and batch- based in-memory workflows (e.g., machine-learning workloads). The broker component implements a set of private interfaces to direct requests from the public APIs to the adequate services.

Spokes can be connected to the VDC network either through a FIONA-based node or through a regular Internet connection, hosting the required APIs on any server or virtual machine. The capability allows to connect cloud services to the VDC, e.g. by storing datasets or offloading computation to a public cloud such as Amazon Web Services or Microsoft Azure. This particular feature makes the VDC easily extensible with little investment: new institutions can join with a single FIONA node and external repositories and cloud resources simply need to implement a small set of APIs.



## Networking

The VDC stitches together the two regional networks — NJEdge and KINBER — to connect Data Hubs at RU-NB, PSU, and Spokes at RU-Camden, RU-Newark, Temple University, Drexel University, and University of Pittsburgh. This regional infrastructure consists of three main components: the Data DMZ backbone, connections to the Data Hubs, and connections to the Data Spokes. It is designed to interoperate with existing regional and national networks, allowing the VDC facility to be easily accessible and expandable. Each network component is described below.

### Data DMZ

A Data DMZ connects the VDC Data Hubs and Spokes. The Data DMZ is a virtual private LAN service (VPLS) configured across existing regional and campus networks. The Data DMZ backbone leverages existing network infrastructures in New Jersey and Pennsylvania. NJEdge, RU, and KINBER have existing resources in 401 North Broad Street in Philadelphia, a popular colocation facility in the region. The Data DMZ is formed by connecting these resources at 10GE using cross connects within the facility. A VPLS network across these networks segregates traffic from and between the Data Hubs and Spokes from the other data within the regional networks.

### Integration with Local, Regional and National Cyberinfrastructure Programs

In addition, the infrastructure allows routed access to organizations connected directly to the statewide NJEdge or KINBER networks. Wide area network access to resources utilized by the VDC projects and external organizations collaborating with the VDC projects is provided through existing connections to regional (OARNet, WVNet) and national (Internet2, ESnet) networks. The Data Hubs connect to the Data DMZ directly or through existing regional network providers (NJEdge in New Jersey/KINBER in Pennsylvania). At both RU and PSU, a FIONA Data Transfer Node [13] following the ESnet specification is connected to their existing science DMZ and then to a 10GE connection. Both Data Hubs connect into the VPLS network, allowing seamless connectivity to other entities. The Data Spokes in each region will connect to the Data Hubs and the Data DMZ using existing network infrastructure. The RU NB campuses are linked via a 10Gbps core network, upgradable to 80Gbps; the RU Camden and Newark campuses connect to the core network using existing 10GE or 1GE infrastructures. Similarly, Drexel and the University of Pittsburgh will use existing 10Gbps connections to KINBER to connect to the PA Data Hub at PSU. As with the Data DMZ, VPLS or equivalent technology will be used to configure a virtual layer2 switch between the Data Hubs and Spokes, forming a virtual Data DMZ, supporting any-to-any (multipoint) connectivity. FIONA-based node installed at the sites will facilitate data transfers to the data hubs and will provide a mechanism for understanding and optimizing end-to-end network performance.

## DATA SERVICES

The Data Services Layer (DSL) provides a virtual cataloging and querying system that utilizes the logical relationship between data elements to connect researchers to the information they need to increase their research quality and interdisciplinary breadth. The data services layer is an innovation of the VDC project to enable researcher interaction with the data network and each other. Science research has been largely siloed within research domains that rarely penetrate other areas of inquiry. This is changing with emerging collaborative research modalities, such as team science, where methods, tools and data from many disciplines are brought to bear on the strategic problems facing society. Scientists are engaging in interdisciplinary research and can use the VDC collaboration space to discover research in other disciplines, identify the relationship of that research to their own interests, and intuitively collaborate to create transformative and impactful open access research.

The DSL is based on in-depth research with fourteen interdisciplinary scientists and two focus groups with more than 50 graduate students and post-doctoral researchers at Rutgers and Temple University. These researchers frequently search for data outside their own fields. They indicated considerable frustration due to the disambiguation of data from creators and due to a lack of trust in data discovered, not knowing the provenance of the data or the credentials of the data creators. Researchers should be able to evaluate data that exists separately from a peer reviewed article describing its creation and intended use.

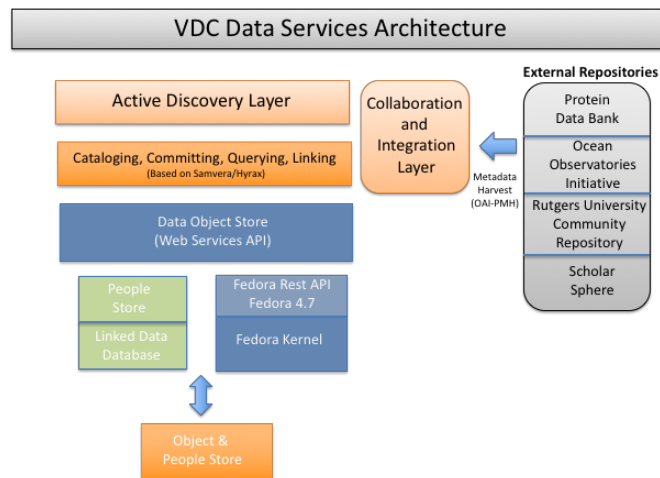


Figure 4. VDC Data Services architecture

To implement the DSL, we define three classes of objects as linked data in order to closely associate people, data and the tools used to analyze that data. This enables us to create links between three broad classes of objects: people, faculty and students who create or deposit the data; data, products of the research; and tools, applications that analyze, visualize, annotate and publish data. In the DSL, we draw an object relation graph using linked data to identify how the various objects relate to each other and the strength of the relationships. We utilize the VDC registration form to capture metadata about people –data depositors and creators– in order to create a metadata record for the creator as linked data that is equivalent to, and can be discovered with, the data itself and ancillary work products (data analysis, visualization, etc.) and the tools used to create or analyze the data. The DSL consists of four main layers illustrated by Figure 4:

1. **Cataloging, Committing and Querying.** In this layer, Samvera 2.3.3 [14] (formerly Hyrax) has been extended to accommodate person and tool as well as resource and collection metadata. Samvera was selected because it is supported by the same robust and active development community that supports the open source repository Fedora 4.7.5. Data is ingested into the DSL as a Fedora object and catalogued as linked data with sufficient context to enable a researcher to determine its relevance for his/her needs. Beyond collecting these resources (data, research products, etc.), the people creating, depositing and using data and tools will also be documented.
2. **Active Discovery Layer.** This layer is the user interface for discovery and access to resources. Linked data is used to create relationships between various data object classes — creators (“people”), datasets, work products, and tools. This layer insures that the necessary relationships are defined, captured, and persisted as data is added to the VDC.
3. **Data Object Store.** This layer contains all the virtual representations of data objects (people, resources, tools), relationships between objects, and the resource locators (DOIs, ORCIDs). The object store includes the Fedora repository for storing data objects and a database for the linked data person directory.
4. **Collaboration and Integration Layer.** This layer leverages the innovative linked data discovery services and includes the core collaboration capabilities, the management of personal collections, setting up project teams, and interfacing to open source tools for analysis, visualization, and annotation.

We discovered through interviews that the most significant commonality across all audiences for selecting and trusting data was the creator—their institutional affiliation and role, whether they could be contacted for a conversation about the data. Directory applications that provide contact information links to publications by researchers are increasingly common. The ORCID research identifier registry provides the opportunity to link publications to the researcher profile. VIVO is a popular ontology-based directory with considerable traction in the United States linking researchers and publications [15]. Others, such as ResearchGate [20], mine the Web for publications of registered researchers and provide email-based prompts to identify and link publications. The linkages between directories and repositories is thus not entirely seamless in most applications but is instead a one-off providing at most a link to finding out more about the data creator. An innovative aspect of the VDC is treating the researcher as metadata for finding, evaluating and selecting data.

The Samvera community spearheaded the development of the Portland Common Data Model (PCDM) [16], a standard for metadata interoperability that includes links between data and their authors. This enables us to seamlessly mine information about data creators, integrate it into the Active Discovery Layer and to it display in the VDC catalog, thus allowing users to select a resource based on authorship. In addition, the data services record data provenance, i.e. links between datasets that are used as input to produce more data. Moreover, the PCDM we developed supports metadata related to software; the model will be deployed as data analysis and visualization tools are added to the VDC, linking data creators, resources and the software used to analyze and create data. This powerful feature will allow researchers to consult the lineage of datasets that are cited in publications or discovered through the Data Services portal.

## EDUCATION AND OUTREACH ACTIVITIES

The VDC presents an opportunity for data collaboration in scientific research and reuse of data which makes it ideal for education. In addition to incorporating the VDC into research-based courses, we aim at incorporating it into general data science/analytics classes so that students can perform large, applied projects in data science and data-intensive research. Rutgers, Penn State, and Drexel have data science and data analytics programs. A stable resource and infrastructure is needed to teach students about working with large datasets, and the VDC offers a realistic test bed for the hands-on projects. Planned education and outreach programs leveraging the VDC platform include learning modules for high school students, undergraduates, graduate students, and early career researchers.

Education and outreach efforts include: *i)* Identification of courses in the data science and computer science curricula that can leverage the VDC to incorporate large datasets and large-scale computing projects at participating institutions; *ii)* A survey of existing resources and tools that has been conducted, and potential partners have been identified; *iii)* The framework has been developed for a Big Data Career Seminar series for undergraduate students that includes: introducing students to career opportunities in data science, data stories on science impact, and application of big data in various industries. The Big Data Career Seminar Series was launched during Spring 2018 at Rutgers University for undergraduate students; *iv)* A high school hands-on workshop, Diving into Big Data using Ocean Observatory Initiative data, which was launched February 2018; and *v)* An Introduction to Data Management Workshop for early career researchers that was developed and delivered at two participating institutions. This seminar addressed issues in data management, stewardship, reproducibility, and curation. Reusability and collaboration are emphasized in the tools that are developed.

Challenges include delivering the educational programs more broadly leveraging the VDC platform for collaboration. Planned activities, in addition to developing online modules, are to curate datasets for use in courses to allow students to learn with real data in previously identified initial courses at Rutgers and PSU. A set of online learning modules will be developed to help students (and faculty) quickly and easily come up to speed with the Collaboratory. A gap analysis will be conducted to identify additional complementary requirements to the existing tools and resources for large scale data management. Our goal is to reuse as much existing material as possible (from sites such as XSEDE, Software Carpentry) and add new material specific to the VDC to ensure everything is usable and researchers and students can use the facility. The learning modules,

tools and resources are available on the VDC website for broad implementation across NJ, PA, and beyond. We will also foster learning communities around using the online modules to enable peer-peer learning. By building peer-learning networks we will be able to build a sustainable data science-learning community and foster increased use of the VDC.

## ENVISIONED BROADER IMPACTS

VDC addresses identified needs by sharing data, analysis tools, workflows that document analysis and results, an integral aspect of data intensive research, education, and innovation. The VDC will make available cyberinfrastructure resources typically not available in the non-intensive and under-resourced campus environment. These will dramatically enhance the quality and reproducibility of data-driven science and researcher productivity. Lessons learned will inform further development of VDC and the development of federated data infrastructures for collaborative data intensive science nationally. VDC will lower barriers of entry for researchers across a broad range of disciplines to data intensive research within specific domains of expertise and collaborative interdisciplinary projects aimed at addressing major scientific or societal challenges.

VDC will enhance substantially the research computing infrastructure for data intensive research on participating campuses and beyond. The intense user engagement in the VDC will ensure a platform that is readily adoptable by scientists, making adoption appealing to any university or consortium. Further, given the expandable and scalable architecture, federation could extend across these structures to realize national data facilities. The extensive efforts required for the first iteration of the VDC and the evolution of the Collaboratory going forward provides opportunities to engage, educate, and train computer science students in myriad aspects of data science techniques and challenges. It also provides a model for engaging relevant departments within universities, like the library, which is critical for successful academic research management.

## CONCLUSION

The Virtual Data Collaboratory is an NSF-funded cyberinfrastructure that can be accessed by researchers, educators, and entrepreneurs across institutional and geographic boundaries, fostering community engagement and accelerating interdisciplinary research. Eventually, the VDC will support end-to-end data intensive scientific applications by federating existing compute and storage resources in New Jersey and Pennsylvania, across seven universities. Its extensible architecture allows other regional schools served by the New Jersey and Pennsylvania high-speed networks to participate and makes it suitable to a large number of science engineering domains.

The project also develops a custom site federation and data services layer for data linking, searching, and sharing, coupling to computation, analytics, and visualization, mechanisms to attach unique Digital Object Identifiers (DOIs), archive data, and broadly publish wider audiences. Its long-term data lifecycle management system will ensure immutable and authentic data and reproducible research. Connecting the system to existing research data repositories, such as the Ocean Observatories Initiative and Protein Data Bank and integrating it into both graduate and undergraduate programs across several institutions, will participate in training the future generation of data scientists to cutting edge practices and cyberinfrastructures.

The end product will be based largely on commercial off-the-shelf technology, leveraging the Hadoop File System for scalable and reliable storage and several high-end big data frameworks such as Hadoop MapReduce, Spark and Alluxio. The development of the compute system is based on a prototype-based action plan, focused on use case applications and core VDC capabilities. A prototype compute system has been deployed at Rutgers along with early deployed FIONA-based nodes (baseline configuration). On the network side, the VDC Data DMZ was established by interconnecting New Jersey and Pennsylvania, along with the regional networks. FIONA and perfSONAR platforms were deployed at the sites and initial benchmarking provided a network performance baseline. The VDC Data Services prototype has been deployed and provides capabilities for creating projects, depositing datasets, searching/browsing, access controls, full text indexing, and account creation. TODO

## ACKNOWLEDGMENTS

This work is supported by NSF grant number 1640834 under the DATANET program. The following people have contributed to this article through their participation in the project: Thu Nguyen, Charles Hedrick, J.J. Villalobos, Mike Scarpellino, James von Oehsen, Ryan Womack (Rutgers University), Wayne Figurelle, Chuck Gilbert, Ryan Gilmore, Kenneth Miller II, Karen Estlund, Sara Peterson, Robert K. Oldendorf (Pennsylvania State University), Edward Chapel, James Stankiewicz, Adam Bathiard (NJEdge), Wendy Huntoon, Michael Carey (KINBER) and Annie Johnson (Temple University).

## REFERENCES

- [1] V. G. Honavar, M. D. Hill, and K. Yelick, “Accelerating Science: A Computing Research Agenda,” *arXiv:1604.02006 [cs]*, Apr. 2016.
- [2] “NSF Facilities Cyberinfrastructure Workshop.” [Online]. Available: <https://doi.org/10.7278/S5SN074P>. [Accessed: 24-Oct-2018].
- [3] W. K. Michener *et al.*, “Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences,” *Ecological Informatics*, vol. 11, pp. 5–15, Sep. 2012.
- [4] M. Lowman, C. D’Avanzo, and C. Brewer, “A National Ecological Network for Research and Education,” *Science*, vol. 323, no. 5918, pp. 1172–1173, Feb. 2009.
- [5] A. C. Catlin *et al.*, “A Cyberplatform for Sharing Scientific Research Data at DataCenterHub,” *Computing in Science & Engineering*, vol. 20, no. 3, pp. 49–70, May 2018.
- [6] S. Wang, H. Hu, T. Lin, Y. Liu, A. Padmanabhan, and K. Soltani, “CyberGIS for Data-intensive Knowledge Discovery,” *SIGSPATIAL Special*, vol. 6, no. 2, pp. 26–33, Mar. 2015.
- [7] A. Chourasia, D. Nadeau, and M. Norman, “SeedMe: Data Sharing Building Blocks,” in *Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact*, New York, NY, USA, 2017, pp. 69:1–69:1.
- [8] L. Smarr *et al.*, “The Pacific Research Platform: Making High-Speed Networking a Reality for the Scientist,” in *Proceedings of the Practice and Experience on Advanced Research Computing*, New York, NY, USA, 2018, pp. 29:1–29:8.
- [9] “Aristotle Cloud Federation.” [Online]. Available: <https://federatedcloud.org/>. [Accessed: 03-Oct-2018].
- [10] “Open Science Grid.” [Online]. Available: <https://www.opensciencegrid.org/>. [Accessed: 02-Apr-2018].
- [11] L. M. Smith *et al.*, “The Ocean Observatories Initiative,” *Oceanography*, vol. 31, no. 1, pp. 16–35, 2018.
- [12] H. M. Berman *et al.*, “The Protein Data Bank and the challenge of structural genomics,” *Nature Structural & Molecular Biology*, vol. 7, no. 11s, pp. 957–959, Nov. 2000.
- [13] “FIONA – Flash I/O Network Appliance.” [Online]. Available: <https://fasterdata.es.net/science-dmz/DTN/fiona-flash-i-o-network-appliance/>. [Accessed: 02-Apr-2018].
- [14] “Samvera.” [Online]. Available: <http://samvera.org/>. [Accessed: 15-Oct-2018].
- [15] “VIVO.” [Online]. Available: <https://duraspace.org/vivo/>. [Accessed: 15-Oct-2018].
- [16] K. Estlund, M. A. Matienzo, D. Fleming, and J. Stroop, “Portland Common Data Model (PCDM): Creating and Sharing Complex Digital Objects,” Dec. 2015.