Council of Data Facilities (CDF) “Shared Infrastructure” Initiative

White Paper

Convened July 24-45, 2019 at the Westin Hotel, Denver Airport, Denver CO
Report completed November 11, 2019
Authored by Tim Ahern, Danie Kinkade, Kerstin Lehnert, Erin Robinson and Lynne Schreiber

Executive Summary

Digital data repositories provide valuable and trustworthy services to the research community, including trusted access to data assets, software tools, services, and data documentation, all of which support transparency and reproducibility of the scientific process. Although geoscience domain repositories have evolved to serve their own specific communities, the growing need for open, discoverable, well-documented cross-domain data discovery, mining, and analysis to answer complex Earth system questions is driving increased repository coordination.

The Council of Data Facilities (CDF) was founded in 2014 to provide a forum for collaboration, coordination, and innovation among Geoscience domain repositories enabling integrative science while at the same time increasing repository productivity through the promotion of reuse of services and adoption of best practices. Since its inception, the CDF has engaged in a number of activities aimed toward aligning member repository best practices, including collaborating with publishers for enabling FAIR principles, supporting Core Trust Seal certification, and leading the implementation of the schema.org approach to exposing facilities and data products at pilot data facilities. The CDF mission also explicitly includes a function to: *Identify and support the development and utilization of shared infrastructure services, including computing services, professional staff development and training services, and related activities.*

Since 2018, the CDF has pursued the establishment of a shared infrastructure environment for digital data repositories, with the vision to co-locate the computational and storage resources with “Cloud” and High-Performance Computing (HPC) environments, which will help address many of the computational and storage needs of CDF members. Although non-trivial, operating diverse domain repositories in the same environment will also naturally lead to the development of methods and procedures for data integration. An important goal of the shared infrastructure effort is to develop best practices, promote the use of interdisciplinary standards in addition to domain standards, and standardize the procedures through which data are discovered, accessed, and used.

Specifically related to this topic is that individual workflows that require data transfer across the Internet at the time of computation will likely be of low performance. By having all data available at a shared infrastructure location, interdisciplinary workflows should be far more flexible and optimized in terms of processing efficiencies. One motivating factor of the shared infrastructure effort is to enable the research community with the tools it needs to easily integrate data from multiple domains and generate interdisciplinary products.

Two EarthCube NSF funded projects, the Alliance Testbed Project (ATP) and GeoSciCloud, investigated elements of a shared infrastructure from 2015 to 2019. The ATP, led by the Interdisciplinary Earth Data Alliance (IEDA), explored a partnership among smaller data communities to share infrastructure and common data services for trusted data curation, while providing domain-specific services for data capture, data access, and data management to their respective communities. ATP developed prototypes of shared data services such as the Data Submission Hub and the IEDA Integrated Catalog, but found that the development of shared

infrastructure required a more scalable, cloud-based solution, and that a more comprehensive initiative at the level of the CDF and involvement of dedicated cyberinfrastructure facilities such as XSEDE would be a more feasible approach.

In GeoSciCloud, two larger CDF data centers, IRIS and UNAVCO, investigated how to operate infrastructures in a variety of cloud environments (e.g. Amazon Web Services (AWS) and Extreme Science and Engineering DiscoveryEnvironment (XSEDE)). Independently, IRIS and UNAVCO found that while commercial cloud infrastructures were more mature, the costs of operating in the AWS environment, exceeded what could be expected for an NSF supported data facility. GeoSciCloud also confirmed that the XSEDE Cloud services built upon on Jetstream and Wrangler approaches were viable.

Based upon the insights from the GeoSciCloud project and the ATP, the CDF initiated a more comprehensive effort in 2018, the Shared Infrastructure Pilot (SIP). The SIP Working Group that included representatives from 12 NSF-funded repositories and two NSF-funded cyberinfrastructure services, embarked on identification of shared infrastructure needs, including computing services, data storage, cyber-security, professional staff development and training, and related activities. The SIP Working Group organized the Shared Infrastructure Workshop in July 2019, with participation by representatives of Geoscience data facilities and the XSEDE partners Indiana University (IU) and the Texas Advanced Computing Center (TACC). We believe that as the shared infrastructure effort moves forward, much of the computational and storage capacity can be met within XSEDE.

Workshop attendees converged on a common understanding of a shared infrastructure, which would not merge facilities, but rather allow repositories to retain autonomy while benefiting from collective infrastructure. They agreed to propose a pilot project that has the following goals:

- *Improve capacity and capabilities*- Develop common shared sustained infrastructure to support all geoscience domain repositories
- *Coordinate and converge on technology and operational procedures*- Increase shared technical understanding, and capabilities by service providers and service users (e.g., NEON, UNIDATA and UNAVCO plan to have a training on Kubernetes in October 2019).
- *Enhance sustainability and resiliency*- Enhance the ability of facilities to provide trustworthy services within the broader community such as curation/accuracy of replicated stored data.

The initial pilot will have a narrow scope and will focus on the most pressing technical needs of computational hardware and shared services and software. If this is successful at a small scale, we would hope that this model may expand beyond our initial 12 GEO and BIO partners.

Five immediate next steps were identified to continue development of the SIP: (1) communicate Shared Infrastructure Workshop output to NSF; (2) work closely with NSF to identify appropriate funding vehicles; (3) incorporate Workshop output (e.g., a draft roadmap) into a more formal proposal; (4) define and agree on the pilot scope, including commonalities and principles as well as to quantify the community needs and costs; and (5) agree on a sustainable approach, leveraging lightweight governance for collective decision making.

The SIP envisions collaboration among three stakeholders - the repositories, the researchers, and NSF. The CDF SIP welcomes early NSF program manager involvement across GEO, OAC, BIO, ENG, and other relevant parts of NSF.

Introduction and Background

The Council of Data Facilities (CDF) was founded in 2014 to provide a forum for collaboration, coordination, and innovation among Geoscience domain repositories that helps to enable integrative science while at the same time increasing repository productivity through the promotion of reuse of services and adoption of best practices. Over the past 5 years, the CDF has engaged in a number of activities aimed toward aligning member repository best practices, such as, collaborating with publishers for enabling FAIR principles, supporting Core Trust Seal certification and leading the implementation of the schema.org approach to exposing facilities and data products with pilot data facilities.

The CDF mission explicitly includes: *Identify and support the development and utilization of shared infrastructure services, including computing services, professional staff development and training services, and related activities*. In 2018, the CDF established a Working Group for "Shared Infrastructure" to 1) conduct surveys to better understand the range, needs, and requirements of CDF members for Shared Infrastructure, and b) to converge on a common vision and roadmap for such Shared Infrastructure.

Motivation

Data facilities in the Earth and Space sciences deliver indispensable services to the science community ensuring discovery, access, reusability, attribution, and preservation of data as essential products and resources of scientific research. Most, if not all, data facilities these days are challenged with the rapidly expanding volume of data that they need to manage, requiring new levels of storage and network capacity and, consequently, resources and team expertise; with the need to comply with international guidelines for trustworthy operation of services, including transparent and standards-based data curation procedures, licensing, risk management, security, and sustainability; with continuously evolving opportunities for enhancing machine access and interoperability of data holdings; and last but not least with stagnant or declining budgets. Establishing shared infrastructure services can potentially help data facilities better address these challenges, making operations more efficient and sustainable, while also leading to better alignment of policies and procedures, enabling machine learning opportunities on a common platform, and opening opportunities for joint developments and innovation to meet future needs.

Experiences from EarthCube Projects

Two EarthCube NSF funded projects, the Alliance Testbed Project (ATP) [EarthCube Grant 1541022] and GeoSciCloud [ICER 1639719] investigated elements of a shared infrastructure

between 2015 and 2019. The ATP, led by the Interdisciplinary Earth Data Alliance (IEDA), explored a partnership among smaller data communities to share infrastructure and common data services for trusted data curation such as DOI registration, Long-Term Archiving, and data submission and access, while providing domain-specific services for data capture, data access, and data management to their respective communities. While ATP developed prototypes of shared data services such as the Data Submission Hub and the IEDA Integrated Catalog, it found that the development of shared cyberinfrastructure (CI) required a more scalable, cloud-based solution, and that a more comprehensive initiative at the level of the CDF and involvement of dedicated CI facilities such as XSEDE would be a more feasible approach.

In GeoSciCloud, two larger CDF data centers, IRIS and UNAVCO, investigated how to operate infrastructures in a variety of cloud environments (e.g. Amazon Web Services (AWS) and Extreme Science and Engineering Discovery Environment (XSEDE). Independently, IRIS and UNAVCO found that while commercial cloud infrastructures were more mature, the costs of operating in the AWS environment, exceeded what could be expected for an NSF supported data facility. GeoSciCloud also confirmed that the XSEDE Cloud services built upon on Jetstream and Wranger approaches were viable. As such XSEDE partners Indiana University (IU) and the Texas Advanced Computing Center (TACC) were invited to participate in the shared infrastructure workshop. As the Shared Infrastructure effort moves forward, we believe much of the computational and storage capacity can be met within XSEDE. While AWS if far more mature than XSEDE's cloud at this time, pricing is a huge concern and deemed high risk to assume that costs for AWS could be contained for the longer term.

Based upon successful NSF-Funded efforts, (e.g. GeoSciCloud and ATP) a more comprehensive CDF effort was initiated in 2018. The Shared Infrastructure Pilot (SIP), composed of the CDF Shared Infrastructure Working Group includes representatives from 12 NSF-funded repositories and two NSF-funded cyberinfrastructure services. SIP embarked on the development and utilization of shared infrastructure, including computing services, data storage, cyber-security, professional staff development and training, and related activities.

Common Data Repository Needs: Initial Steps

During the summer 2018 CDF assembly, CDF members began building consensus around the concept of sharing cyberinfrastructure. Common needs were expressed in the areas of hardware and computing resources, software licensing, and training. By identifying shared needs of CDF members, data centers will have the opportunity to become more efficient, developing domain specific infrastructure when needed, but relying on shared infrastructure whenever possible. In late 2018 a working group was formed to further explore this idea. Three surveys and several breakout sessions were used to understand the common data repository needs.

The initial survey, solicited input from all members of the CDF and informed the SIP that the greatest interest in shared infrastructure related to the hardware (processors and storage resources) and software frameworks (e.g. virtualization, operating systems, database management systems, deployment methods, etc.) There was also significant interest in training in technologies and best practices across the CDF respondents. While there was interest in benefits to be gained in shared licensing (e.g. minting of DOIs and ORCIDs for instance) it was at a much lower importance. Appendix C

A survey taken just before and updated during the Shared Infrastructure Workshop focussed on the two highest priority areas 1) shared hardware and software infrastructures (Appendix D) and 2) training opportunities (Appendix E).

Appendix D shows that the current and anticipated hardware needs of the CDF centers were modest and fell within the capacities of NSF supported activities such as XSEDE. The biggest disconnect between the cloud opportunities in XSEDE and the needs of the CDF was the allocation model. CDF data centers require very high up time to meet the needs of their various communities. There was also a resonance in the types of frameworks many of the CDF members were using although there was a great disparity between the level of maturity at the various centers in using those frameworks.

Appendix E identifies a variety of topics that received attention related to training workshops that the SIP showed interest in. There was enthusiasm to establish training workshops as low-hanging fruit and an opportunity to make significant and important progress to improve CDF collaborations..

While there was a presentation from ORCID at the workshop, there was little interest in setting shared licensing as a high priority. Governance was also briefly discussed but no specific actions or next steps were identified at the workshop.

The takeaway message from the workshop in terms of common needs of the SIP was that hardware infrastructure, software frameworks and applications, and establishing training workshops on a variety of topics should be the focus of the SIP in its early phases.

A final outcome that received broad support at the workshop was that this provides an opportunity to improve communications across the CDF membership.

Value proposition

Earthcube's primary goal since its inception has been focussed on solving the more complex, whole earth scientific problems that lie outside a single domain. For historical reasons each domain supported by NSF, as well as other federal agencies, has built its own domain specific

infrastructure usually in a geographically different location. The ability to integrate data across domains in conjunction with isolated repositories is not optimal to support interdisciplinary research.

While various efforts within EarthCube (e.g. P418, P418GUI, P419) have made cross-disciplinary discovery a more realizable goal, there remain difficulties when integrating data from distributed centers related to

- different domain vocabularies,
- different approaches to access and use data
- Multiple and different formats, and
- Integrating information with Big Data characteristics at a single location where integrated products can be formed is a challenge.[1] In this context, Big Data possess one or more of these characteristics:
    a. **Volume:** The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not.
    b. **Variety:** The type and nature of the data. This helps people who analyze it to effectively use the resulting insight. Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
    c. **Velocity**: In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.
    d. **Veracity**: It is the extended definition for big data, which refers to the data quality and trustworthiness.

By developing a shared infrastructure environment where computational and storage resources are co-located with “Cloud” and HPC environments, some Big Data problems will become more tractable. Also, by having these resources available in the same environment there will be a natural tendency to develop methods and procedures that ease integration. An important goal of the Shared Infrastructure effort is to develop best practices, promote the use of interdisciplinary and domain standards, and align procedures through which data are discovered, accessed and used.

Specifically related to this topic is that individual workflows requiring data transfer across the internet at the time of computation will often be of low performance. By having all data available at a shared infrastructure location, workflows should be far more flexible and optimized in terms of processing efficiencies. One motivating factor of the Shared Infrastructure effort is to enable the research community with the tools it needs to easily integrate data from multiple domains and produce interdisciplinary products.

[1] https://en.wikipedia.org/wiki/Big_data

The group brainstormed value propositions of a shared infrastructure activity from the perspectives of both repositories and research scientists. (Appendix F)

- Enabling interdisciplinary science
- Connection to big ideas
- Democratizing otherwise achievable resources
- Leveraging lessons learned and expertise
- Increased resiliency

Specifically, the shared infrastructure pilot group:

- Will help **research scientists do interdisciplinary & innovative science** by easily accessing data and services across the partners and using big data and HPC approaches,
- Will help **data facilities focus on domain specific curation and services** by sharing infrastructure such as storage, knowledge, and expertise that is common across all repositories/facilities, becoming more consistent
- Will help **NSF accelerate scientific discovery** by supporting data resources being located in close proximity but curated and managed by sustainable domain repositories.
- Will accelerate interdisciplinary scientific discovery by supporting more effective domain curation of NSF supported scientific results, while reducing researcher effort in data discovery, access and use. The sharing of repository infrastructure has the potential to support natural alignment of practices and technologies in addition to facilitating the integration of data-related elements such as data and metadata formats, vocabularies, etc.

And of great importance to NSF, sharing computational and storage resources should reduce the funding by NSF to support multiple domain repositories.

Significant Outcomes

At the Shared infrastructure Workshop attendees converged on a common understanding of a shared infrastructure which would not merge facilities, but rather allow repositories to retain autonomy while benefiting from collective infrastructure. The goals of the proposed pilot are to:

- *Improve capacity & capabilities* - Develop common shared sustained infrastructure to support all geoscience and other NSF domain repositories
- *Coordinate and converge on technology and operational procedures* - Increase shared technical understanding, and capabilities by service providers and service users. NEON, UNIDATA and UNAVCO plan to have a training on Kubernetes in October 2019. A complete list of training needs are found in Appendix E.

- *Enhance sustainability and resiliency* - Enhance abilities of facilities to provide trustworthy services within broader community curation/accuracy of replicated stored data
- Created a cohort for a Shared Infrastructure Pilot (SIP)

Data facilities realize the benefits of sharing infrastructure

- To improve capacity & capabilities
- To coordinate & converge on technology and operational procedures
- To enhance sustainability and resiliency
- To potentially reduce infrastructure costs

The CDF Shared Infrastructure Initiative is defining and understanding the critical components and challenges in implementing and operating shared infrastructure.

- Technical, organizational, social

A Napkin Drawing activity yielded high overlap in conceptual understanding of a possible shared infrastructure environment. Appendix G

Identified Gaps

Better identification of needs of the Shared Infrastructure partners (survey and identified workshops and trainings).

Survey the landscape of other XSEDE type resources and services options beyond TACC
Different data facilities are at different stages of their funding cycles and ability to use shared services, so will be an ongoing activity before the entire CDF cohort will be able to participate. We believe that individual CDF members will join the shared infrastructure when joining works best for them.

Goals

The roadmapping exercise on the first day yielded two sets of goals and milestones; one each for technical and programmatic themed activities. Technical participants articulated goals heavily focused on infrastructure platforms, data storage, and core services. Additional goals were more intangible and focused on accepting a process of trial and error, a desire to drive down cost and recover time to focus on domain specialization, and achieving resilient and responsive systems in the face of disasters.

Programmatic goals were articulated as visions of a desired end state but had overlap with technical goals with respect to resiliency in repository operations and their data assets.

Additional programmatic goals included:

- increased shared technical understanding between service providers and service users
- enhanced capabilities of facilities to provide trustworthy services within broader community curation and trustworthiness of replicated stored data
- cross-training of knowledgeable repository workforce to achieve resilient and effective repository operations in support of their domain communities
- common shared sustained infrastructure to support all subdomain repositories
- increased collaboration between repositories and science by bringing science to the data (i.e., leveraging jupyter notebooks and similar distributed processing engines), thus the need to download data will be eliminated

Pilot Roadmap

In addition to goals and vision, each breakout group was asked to develop a roadmap including milestones to achieve the goals. Milestones from both groups were merged during the report out and begin at a pre-funding point in time, to years 1-5 of a potential funded pilot project. (Appendix H)

Pre-funding Activities:

- In depth analysis of scope
    - Develop commonalities & principles
    - Identify & quantify community needs
    - Pilot projects leveraging multiple SIP partners
- Define Sustainability Approach
    - Governance
    - Funding options

Note: By the end of each year of funding, some facilities will have achieved these milestones while others may not. Partners will move at different rates that match their capabilities and resources.

Year 1:

- Recovery, backup and replication of data
- Utilize shared storage for SIP partners, moving data assets into XSEDE on a test basis
- Work with XSEDE partners to marshall sufficient VMs and processors to support initial services
- Assessment of a pilot project to see how things are working and address any scaling needs
- All pilot members register organization, data sets, software and services in P418 schema.org (enable search space and time). Motivate P418 to support streams and not just file-based products
- Training technical staff at facilities.

- Work with TACC & Indiana University to fund XSEDE in a Pilot including a new allocation method

Year 2:
- All assets migrated
- Deploy data center service stack for operation
- Develop service level expectations including uptime of services and products
- Production ready for some services
- Begin metrics development

Year 3:
- Production versions of most services
- Monitoring of metrics

Year 4:
- Most pilot facilities have migrated most key infrastructure
- Production for almost all services
- Continuous integration
- Identify gaps
- Identify new needed services
- Update metrics based upon experience to date

Year 5:
- bridge gaps
- Begin to Develop cross disciplinary work flows

Recommendations and Proposed Next Steps

The initial pilot will have a narrow scope and will focus on the most pressing technical needs first of computational hardware and shared services and software. If this is successful at a small scale, we would hope that this model may expand beyond our initial 12 GEO and BIO partners.

Five immediate next steps were identified to continue development of the SIP:
1. Communicate Shared Infrastructure Workshop output to NSF;
2. Work closely with NSF to identify appropriate funding vehicles;
3. Incorporate Workshop output (e.g., a draft roadmap) into a more formal proposal;
4. Identify the initial pilot scope, including commonalities and principles as well as to quantify the community needs and costs; and
5. Agree on a sustainable approach and an initial lightweight governance for collective decision making.

The SIP envisions collaboration among three stakeholders - 1) the repositories, 2) the researchers through existing repository linkages, and 3) NSF. The CDF SIP welcomes early NSF program manager involvement across GEO, OAC, BIO, ENG and other relevant parts of NSF.

Further refinement of specific infrastructure needs

Through continued engagement of the SIP

Host Webinars:

- Bonnie Hurwitz for different resources outside TACC
- NASA DAAC's for lessons learned (or from other federal agencies who have shared infrastructure)

There was considerable enthusiasm related to shared training workshops for CDF members. Participants identified workshops that they would be able to act as trainers as well as workshops that they viewed would be useful for them to attend. Appendix E. lists workshops that were identified by the SIP partners that would be useful.

Participants were unanimously in favor of continuing engagement in the shared infrastructure pilot activity. It was clear that several of the SIPs were enthusiastic participants of a shared infrastructure and they might lead the involvement and others that are less ready to benefit from a shared infrastructure at this time would follow. We anticipate that this will be a phased approach and repositories will determine when they are ready to move to a shared infrastructure.

Acknowledgements

This Workshop and Workshop report were supported by a National Science Foundation EarthCube Grant 1541022 to K. Lehnert. We gratefully acknowledge the logistical services of Lamont-Doherty Earth Observatory and the EarthCube Science Support Office, and facilitation services of the Earth and Space Information Partners (ESIP).

ICER 1639719 T. Ahern
EarthCube Building Blocks: Collaborative Proposal: Deploying Multi-Facility Cyberinfrastructure in Commercial and Private Cloud-based Systems

OCE-1435578 D. Kinkade
Biological and Chemical Oceanography Data Management Office (BCO-DMO): A System for Access to Ecological and Biogeochemical Ocean Data

Appendix A. Participant List

Workshop Planning Committee:

Tim Ahern Incorporated Research Institutions for Seismology (IRIS) tim@iris.washington.edu
Danie Kinkade Biological and Chemical Oceanography Data Management Office (BCO-DMO) dkinkade@whoi.edu
Kerstin Lehnert Integrated Earth Data Applications (IEDA) lehnert@ldeo.columbia.edu
Erin Robinson Earth Science Information Partners (ESIP) erinrobinson@esipfed.org
Lynne Schreiber EarthCube Science Support Office (ESSO) lschreib@ucar.edu

Attendees:

Tim Ahern Incorporated Research Institutions for Seismology (IRIS) tim@iris.washington.edu
Suzanne Carbotte Rolling Deck to Repository Program (R2R) carbotte@ldeo.columbia.edu
Jerry Carter Incorporated Research Institutions for Seismology (IRIS) jerry@iris.washington.edu
Dru Clark Geologic Data Center (GDC) dclark@ucsd.edu
Ethan Davis UNIDATA edavis@ucar.edu
Maria Esteva Digital Rocks Portal maria@tacc.utexas.edu
Vicki Ferrini Integrated Earth Data Applications (IEDA) ferrini@ldeo.columbia.edu
Niall Gaffney Texas Advanced Computing Center (TACC) ngaffney@tacc.utexas.edu
Corinna Gries Environmental Data Initiative (EDI) cgries@wisc.edu
David Hancock Indiana University (IU) dyhancoc@iu.edu
Steve Jacobs National Ecological Observatory Network (NEON) sjacobs@battelleecology.org
Danie Kinkade Biological and Chemical Oceanography Data Management Office (BCO-DMO) dkinkade@whoi.edu
Christine Laney National Ecological Observatory Network (NEON) claney@battelleecology.org
Kerstin Lehnert Integrated Earth Data Applications (IEDA) lehnert@ldeo.columbia.edu
Chuck Meertens UNAVCO chuckm@unavco.org
Eric Olson ORCiD e.olson@orcid.org
David Philips UNAVCO dap@unavco.org
Mohan Ramamurthy UNIDATA mohan@ucar.edu
Erin Robinson Earth Science Information Partners (ESIP) erinrobinson@esipfed.org
Lynne Schreiber EarthCube Science Support Office (ESSO) lschreib@ucar.edu
Mark Servilla Environmental Data Initiative (EDI) mark.servilla@gmail.com
Martin Seul The Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) MSeul@cuahsi.org
Adam Shepherd Biological and Chemical Oceanography Data Management Office (BCO-DMO) ashepherd@whoi.edu
Karen Stocks Geologic Data Center (GDC) kstocks@ucsd.edu
Alexander Stone Continental Scientific Drilling Coordination Office (CSDCO) alexm@umn.edu
Chad Trabant Incorporated Research Institutions for Seismology (IRIS) chad@iris.washington.edu
Eva Zanzerkia National Science Foundation (NSF) ezanzerk@nsf.gov

Remote Attendees:

Steven Whitmeyer National Science Foundation (NSF) swhitmey@nsf.gov
Doug Fils Ocean Leadership dfils@oceanleadership.org

Appendix B. Workshop Agenda

	Day 1 - Wednesday, July 24	
	Room: Spruce II	
	Wifi password: Lamont2019	
	Remote access: https://global.gotomeeting.com/join/829639373	
	You can also dial in using your phone. +1 (571) 317-3116. Access Code: 829-639-373	
7:45	Breakfast/Registration	
8:00	**Welcome, Session Goals & Overview**	
9:00	Niall Gaffney, TACC	
9:15	David Hancock, IU	
9:45	Eric Olson, ORCID	
10:30	Break	
10:45	Programmatic Needs Conversations	Tech Needs Conversation
11:45	Report back on Needs Conversations	
12:30	Working Lunch	
2:00	Mission	
2:15	Roadmap Introduction	
2:30	Roadmap Breakout - Programmatic	Roadmap Breakout - Tech
4:00	Break	
4:30	Report Back & Bring gov/Roadmap together	
5:30	Last Word from Attendees	
6:00	Meeting ends	
6:30	Reception & Dinner	
	Day 2 - Thursday, July 25	
	Remote access: https://global.gotomeeting.com/join/829639373	
	You can also dial in using your phone. +1 (571) 317-3116. Access Code: 829-639-373	

8:00	Continental Breakfast/Registration Open
8:30	Recap from Day 1 (main presentation deck)
9:00	Conversation on Day 1 Output with NSF Program Officers
9:30	Governance Discussion & White paper outline
10:00	Break

Appendix C. Pre-Workshop Survey Results (April 2019)

A pre-workshop survey was distributed to repositories who expressed interest and commitment in advancing the idea of common or shared infrastructure. The survey results were used to refine workshop scope.

Priorities
12 facilities ranked 1st and 2nd priorities as:
computational hardware -- 75% of the facilities
shared software both at the OS level and general application software -- 58%
cross-center training and development of best practices -- 42%
education and outreach -- 17%
shared licensing (e.g. ORCIDs, DOIs), -- 8%

Statement(s) of Intent for participating in the CDF Shared Infrastructure Workshop and how it relates to individual facility infrastructure needs

- **NEON** produces and processes thousands of heterogeneous data streams from field observations, in situ sensors, and remote sensing. The variety and volume of data requires novel solutions, and we are interested in finding synergies and efficiencies in potential partnerships with other CDF member organizations.
- **CUASHI** intends to participate in the workshop to convey the needs of the hydrologic community to successfully deliver data services to the community. As storage and compute needs increase the challenges for the domain repositories increase as well and requires new approaches to solve the needs. Current approaches using local resources and/or commercial cloud providers are often difficult or cost prohibitive. I participate to learn more about needs and practices at other repositories and develop ideas on how to work in conjunction with other repositories to serve the infrastructure needs of the hydrologic community.
- The **Digital Rocks Portal** is a data lifecycle curation and analysis infrastructure hosted at the Texas Advanced Computing Center at UT Austin. This collaboration has enabled implementing innovative solutions for data curation, enhanced visualization, and data access due to the integration of national shared computational resources. Issues implicit in the Focus Areas of the CDF workshop such as best practices, users training, services evaluation and impact, data and services marketing, funding opportunities, and prioritization of functionalities surround any decision related to shared infrastructure. As DRP seeks to move forward to incorporate Machine Learning, improve batch data and metadata services for scalability, and find a sustainability path, it is key to participate in the discussions of this workshop, and to coordinate activities at the CDF level.
- **BCO-DMO** intends to fully participate in the workshop by sending its Director and Technical Director to inform on the needs of BCO-DMO as a domain repository. At a 100% soft money institution with a 65% overhead rate, purchasing external

cyberinfrastructure (CI), which is taxed at the overhead rate, is cost prohibitive to modern day CI architectures and economies of scale. Current offerings at XSEDE are compelling and helpful, but without a hardened commitment lasting at least as long as our funding cycle (5 yrs) makes it challenging to adopt. This leaves us working with local CI services at an institution with support M-F 8-5 making it difficult to manage and maintain production-quality IT services. Shared infrastructure could be an important component in the success of smaller domain repositories at NSF.

- We are maintaining the **Environmental Data Initiative** on NSF funding and are very aware of the cost and potential efficiencies that could be gained through sharing. We have been part of this conversation for a long time including obtaining funding for and organizing a workshop with a similar subject in 2015. This is a difficult topic and the discussions need to take place at many different levels of which the actual repositories are only one. We are prepared to use shared infrastructure and develop shared practices if efficiencies can be gained without loss in service to our customer.
- The **Geological Data Center** is manages several oceanographic data collections / access portals. We are interested in shared storage and compute options, and cross-center training and development of best practices
- Challenges for our facility [**R2R (Rolling Deck to Repository)**] include handling growing data volumes and new cyber security needs, as well as staying abreast of new developments in data management and cyber infrastructure while at the same time dealing with flat budgets and highly over committed staff. We are very interested in how CDF shared infrastructure and might help us meet these challenges. Cross-center training and development of best practices as well as shared computer resources that could help us address our storage challenges are of particular interest.
- At **UNAVCO** as an NSF-funded Facility, we are continually looking for ways to improve efficiency, robustness, and capabilities of our Geodetic Data Center. With IRIS, we have been investigating greater use of cloud resources (commercial and private) under an EarthCube grant “GeoSciCloud”. The results are promising and encourage us to want to be part of a future shared infrastructure activity that would lead to NSF commitments for longer term cloud computational resources as well as shared knowledge. We plan to share our experiences at the CDF workshop and learn from others.
- **Unidata**, a founding member of the CDF, is a long-standing data facility in the geosciences. I am interested in the discussion on how we can leverage the resources, capabilities, and expertise of the Shared Infrastructure and Services that is envisioned, toward advancing Unidata's and EarthCube's mission.
- The facilities I represent (**CSDCO and LacCore**) are the NSF facilities for continental drilling and coring. We generate, transform, serve, and archive data from scientific drilling and coring projects; develop and maintain software for data visualization, workflow support, and data distribution and access; and collaborate with other community data resources to advance community development and data management priorities.

- **IRIS** has already done performance testing of operating our data center infrastructure in cloud environments through the GeoSciCloud Building Block. Shared infrastructure for the CDF would be of direct benefit to IRIS and it is our preferred method moving forward. We would be active users of shared hardware/software/best practices/shared licenses in such an environment. Our initial results indicate that if the shared environment addresses certain requirements satisfactorily such shared infrastructure would be of tremendous benefit to IRIS and the CDF in general.

Appendix D. In-Workshop Survey of Hardware Infrastructure Needs

Although a pre-workshop survey helped to inform us as to various shared infrastructure needs, a short survey was conducted during the workshop of the actual members of the group of pilot CDF members that expressed interest in and attended the workshop. This focussed on two areas, an indication of the number of processors (cores) and storage requirements (terabytes) needed across CDF members attending the workshop.

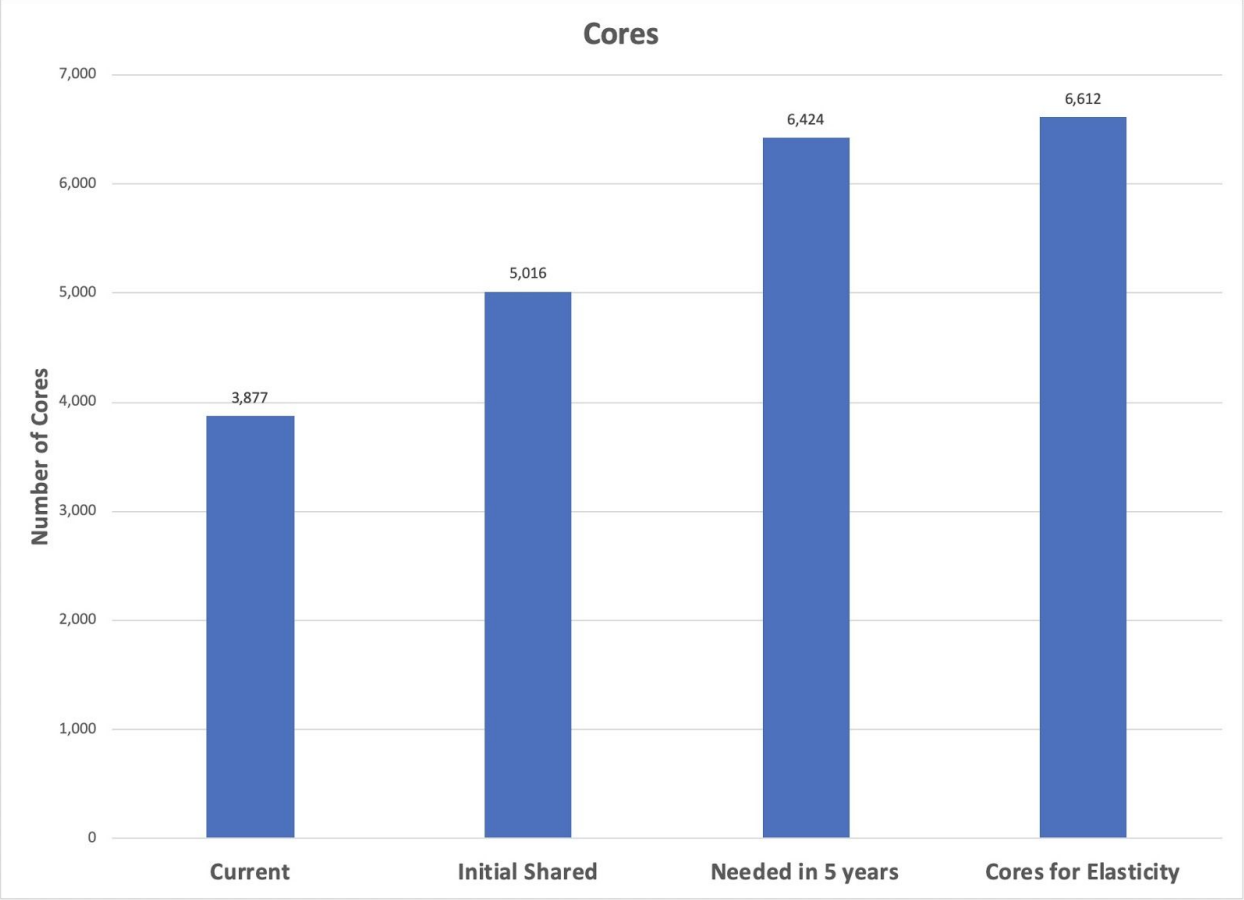


Figure 1. Estimate of Cores Needed. The above figure shows (from left to right) 1) current aggregate number of cores in use at the pilot data centers (3,877 cores), 2) an estimate of the cores initially needed in the shared infrastructure (5,016), 3) estimated cores needed in 5 years (6,424 cores) , and finally 4) cores that might be needed on a short term basis for short term elastic increases in processing power (6,612). For instance shortly after a natural disaster when products must be produced rapidly.

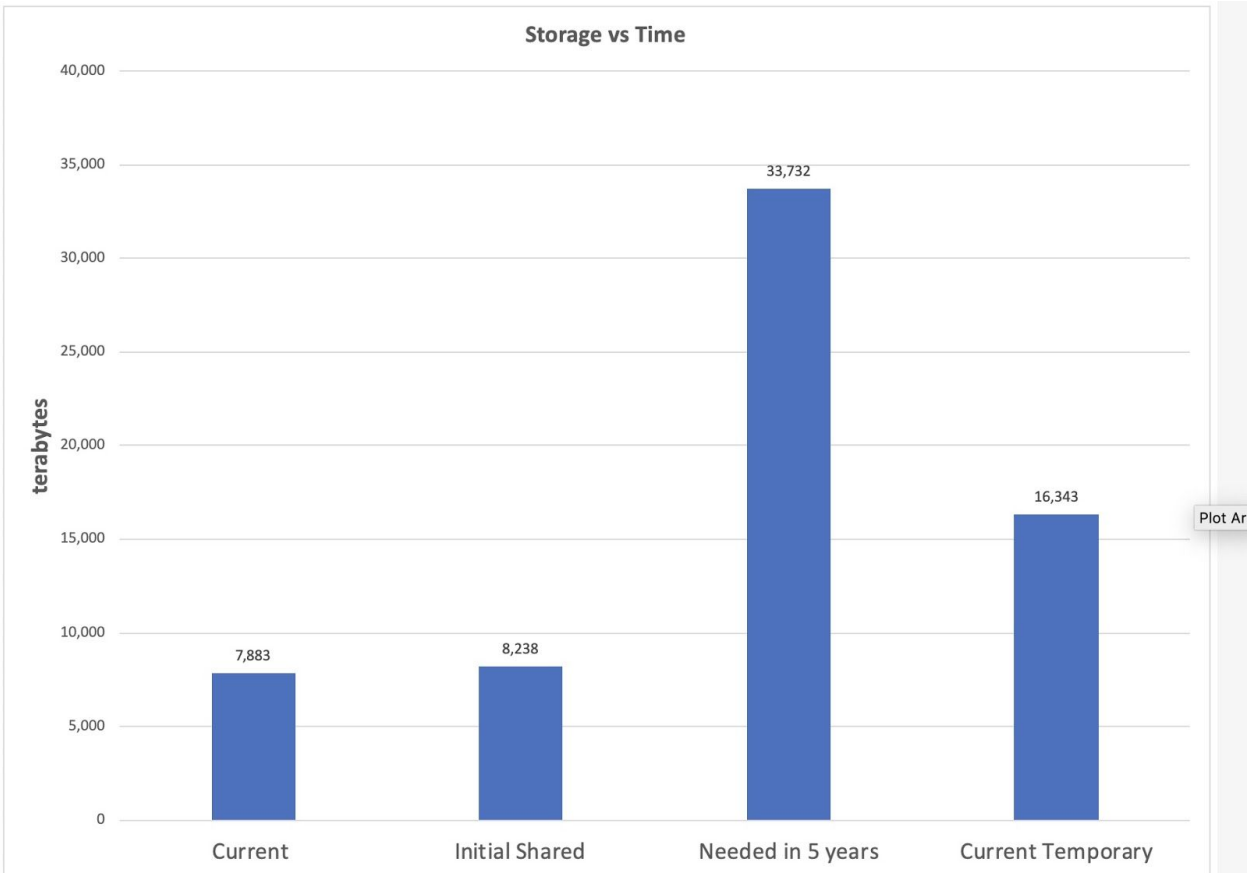


Figure 2. Estimate of Storage Needed. The above figure shows (from left to right) 1) current storage in use at the pilot data centers (7,883 terabytes) , 2) an estimate of the storage initially needed in the shared infrastructure 8,238 terabytes) , 3) estimated storage needed in 5 years (33,732 terabytes), and finally 4) storage that might be needed on a short term 16,343)

Appendix E. Training Workshops that were identified as being useful to the SIPs.

THIS IS JUST A LIST OF WORKSHOP TOPICS MENTIONED BY PARTICIPANTS

- Dealing with persistent data
- Running a kubernetes cluster, ingresses, load balancers, service mesh, tracing, prometheus, etc
- Architects, engineers and other technical staff would learn how to transition to object storage from traditional, file storage systems.
- Learn best practices for deployment in and usage of a shared infrastructure environment. Including general best practices for software development and maintenance that precede deployment to production.
- Learn how to use Function-as-a-Service, aka Servereless, frameworks for efficient and easy deployment of data access web services, data processing services and other workflows.
- Learn how to use Kubernetes to orchestrate service deployment for large and small scale ecosystems with elastic capability.
- Share experience, successes and dead-ends, between centers using common infrastructure.
- lessons learned using S3-compliant object stores
- best practices for containerization, lessons learned
- best practices for designing serverless architectures
- Technical training for technical staff at UNAVCO
- Training for user community
- Best practices for facilities to leverage facility resources
- Lessons Learned
- Clean data, meaningful data package, complete metadata
- We are already using Docker heavily and starting to use Kubernetes but it has been a few individuals learning on their own (with help from Jetstream staff).
- In depth information on performance characteristics, best usage patterns, and considerations.
- We have experimented some with this on AWS but would like to look at moving more heavily in this direction.
- Training on install, configuration, how to manage users and user spaces, how to use Kubernetes to scale to large numbers of users
- We are interested in training in systems issues like deploying containers and kubernetes
- UI-UX design

Appendix F. Value Proposition

The group brainstormed value propositions of a shared infrastructure activity from the perspectives of both repositories and research scientists...

Our shared infrastructure pilot group will help **research scientists**

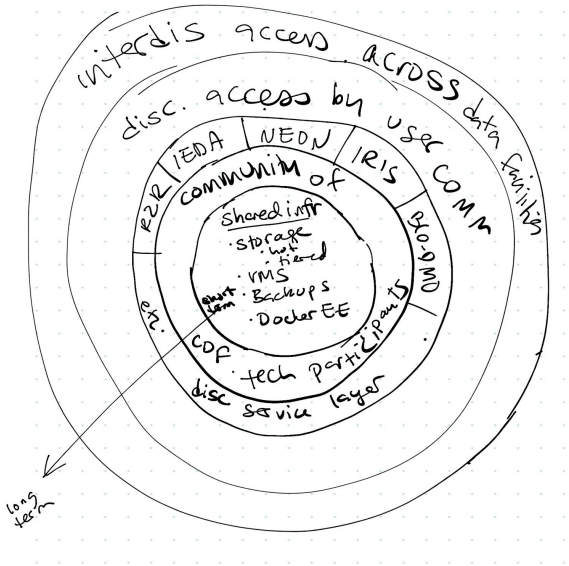
- Do interdisciplinary Science by easing access to data and services and interdisciplinary workflows in general.
- Have easier, faster, and more reliable access to data and tools that over time, become more interoperable and usable and will advance both disciplinary and cross-disciplinary science
- Do innovative science through advanced data analysis by having convenient access to all data and computing resources
- Have more customized, higher level, data products, more quality
- Educate a new generation of scientists with big data
- Support and expand machine learning capability
- Take the scientist to the data, not the data to the scientist
- Access and utilize data resources and more easily use big data and HPC approaches for by taking the scientist to the data

Will help **data facilities**

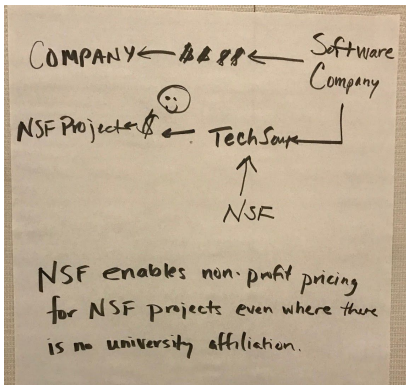
- Focus on the domain expertise to manage and curate their data by off loading the infrastructure services that are common to all domains
- Provide more common data and access layers more easily by leveraging common cyber infrastructure
- Let CI address background mechanical processes (disk maintenance, VMs, disk space, cores, load balancing)
- Focus on domain specific curation and services by sharing infrastructure storage, knowledge, expertise that is common across all repos.

Appendix G. Napkin Drawings

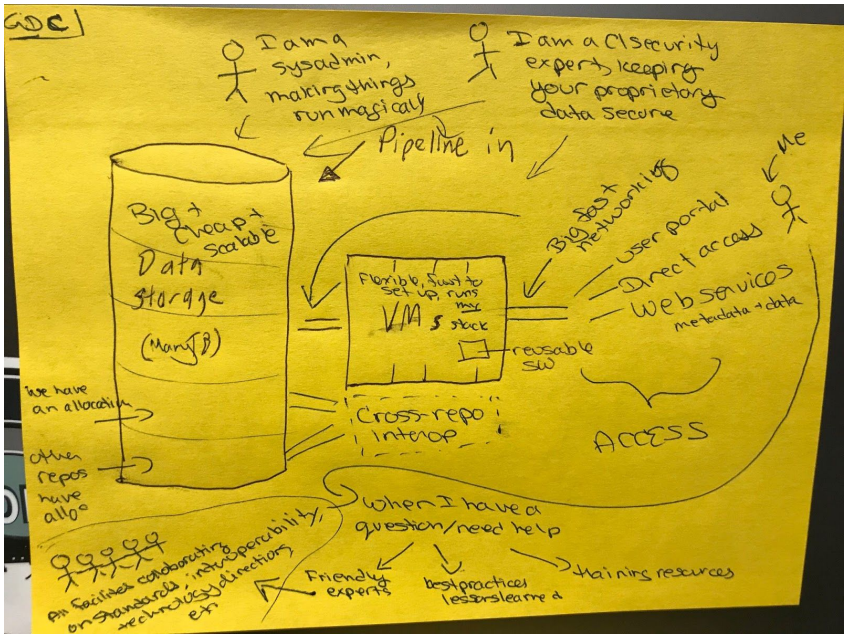
Caption:
With repositories using shared infrastructure (buckets of raw materials) for storage, hosting and compute, a common set of tools and training can all be combined to build an ecosystem where complex questions can be analyzed.



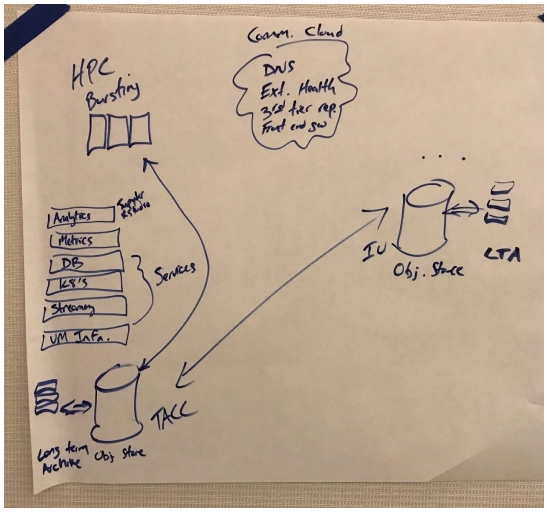
Caption:
Shared infrastructure including storage, VMs, backups and Docker EE unites a community of practice - technical specialists from the CDF who share knowledge and expertise. Disciplinary systems sharing the infrastructure provide access and services to their user communities. By coalescing around shared infrastructure and common approaches we can eventually facilitate access and discovery across disciplinary boundaries.



Caption:
Software licenses can be extremely expensive, especially for non-academic institutions. NSF projects could potentially cost less if NSF could enable non-profit pricing for all work associated with NSF projects, even if the organization doing the work has no university affiliation.



Caption:
Shared storage and compute, supported by experts such as system administrators and security experts, facilitate efficient set-up, use and reuse of components and foster cross-repository interoperability. A community evolves in parallel to promote training, best practices, and peer-to-peer advising.



Caption: Cloud CI complemented by people. Redundant layers of infrastructure replicated between two computing centers (IU & TACC). At the core is object storage with a path to long-term archive. Coupled with the storage is virtual infrastructure with service layers on top for streaming data, kubernetes, database services, metrics (telemetry), and analytics (e.g. JupyterHubs/RStudio). Workflows could burst to HPC services within XSEDE or elsewhere. Commercial cloud services can be leveraged where desired or particularly effective (external DNS, application health monitoring, 3rd tier storage repository, and front-end gateway services)

Appendix H. Pilot Roadmap

Pre-funding Activities:

- In depth analysis of scope
    - Develop commonalities & principles
    - Identify & quantify community needs
- Define Sustainability Approach
    - Governance
    - Funding options

Year 1:

Note by end of year 1 of funding, some facilities will have achieved these milestones while others may not. Partners will move at different rates

- Recovery, backup and replication of data
    - Only a small subset of repositories had fully functioning redundant centers and others only had a cold backup but sometimes in close proximity to the primary copy of data. Shared Infrastructure will better make the data assets secure and less vulnerable to loss.
- Storage
    - As shown in Appendix D, the storage needs of the pilot group is not insignificant and is expected to approach 35 petabytes within 5 years with a short term need of another 20 petabytes on a shorter term basis. This need is better served using a shared infrastructure.
- Sufficient VMs
    - Also shown in Appendix D, the requirement for cores is not excessive estimated currently at about 4,000 cores and increasing by roughly 50% in the next 5 years. For elastic computational needs, it is estimated that the number of cores could reach 6,500 cores just for elastic computational demand. This need is much better realized using shared NSF resources.
- Assessment of Trial to see how things are working and how accurate quantities
    - With the pilot group of data centers in place, this will allow us to more fully estimate actual computational needs and better estimate future needs once the pilot group of centers is operating in the shared infrastructure.
- All pilot members registered organization, data sets, software and services in p418 schema.org (enable search space and time)
- Training of facility's technical staff
    - There was significant interest in establishing training workshops where the centers could share their own knowledge as well as receiving from other centers, Examples of some of the things in which there was interest included are in Appendix F.

Year 2:

- All assets migrated

    - In the second year, participating Pilot members will have migrated all data and selected services to the shared infrastructure
- Deploy data center service stack for operation
    - Key services from participating pilot centers will be deployed in the shared infrastructure,
- Develop service level expectations
    - The CDF pilot data centers will identify the quality of services that participating members are expected to meet.
- Production ready for some services
    - Key services needed by our communities will be ready in a production mode.
- Begin metrics development
    - Initial identification of key metrics that shared data center partners must begin tracking.

Year 3:

- Production stage for most services
- Monitoring of metrics

Year 4:

- Most pilot facilities have migrated most key infrastructure
    - Production for almost all services
- Continuous integration
- Identify gaps
- Identify new needed services
- Update metrics based upon experience to date

Year 5:

- bridge gaps
- Begin to Develop cross disciplinary work flows