Project Name: CyVerse https://cyverse.org

Project Description: CyVerse, headquartered at the University of Arizona, is an NSF-funded, 15-year $115M project that provides cyber-infrastructure (CI) comprising of data storage and management, image analysis, cloud services and more, and training in the use of computational tools to US and international basic research scientists. CyVerse's CI addresses many common barriers associated with data and computation scalability, usability, and extensibility by building upon these proven platforms and technologies and by adopting the best practices from science domains where these challenges have already been addressed. CyVerse's CI provides web accessible tools and well-described application programing interfaces (API) for performing data analysis and managing data driven collaborations, with an emphasis on federating data and consuming computational resources from multiple providers such as NSF funded XSEDE resources (Extreme Science and Engineering Discovery Environment), campus clusters, and commercial cloud providers. CyVerse's CI allows communities to readily utilize a wide array of tools and services, and if required extend the CI to accommodate their specific research community needs. CyVerse provides an avenue for researchers to share their research data, software tools, and analysis pipelines with their collaborators and/or a large community of users without burdening the researchers to provision the underlying computational infrastructure.With nearly 100,000 users worldwide, CyVerse cyberinfrastructure is built for reproducible, computational data analysis and management, teaching and training, and collaborations across diverse disciplines and domains — from life science to human health, agriculture, geology, earth sciences, astronomy, defense, and more. CyVerse is deployed in multiple countries (US, UK, Austria), referenced in over 1200 peer reviewed publications, and trained tens of thousands of students and researchers.

CyVerse's key products are:

- The **Discovery Environment** (DE) data science workbench and its extension **VICE** – web-based platforms to integrate and run ANY app for data management, data analysis, data science, data visualization, data exploration, and data discovery
- **Atmosphere –** CyVerse's on-demand cloud native compute resource
- The **Data Store** and **Data Commons** – 8 Petabytes of secure, high-performance cloud storage from which you can seamlessly manage, share, analyze, and publish final data products with DOIs.
- **Science APIs –** giving you access to CyVerse's backend infrastructure services for building custom projects and cyberinfrastructure
- **Enhanced security** – features available for sensitive and restricted data including ITAR and HIPAA compliance
- **Bisque**: Image analysis and management with the ability to handle nearly 300 image formats for 2D, 3D, 4D, multispectral data, and associated metadata
- **Learning Center**: Tutorials, webinars, workshops, tours, and documentation for all CyVerse services

Contact: Dr. Eric Lyons (ericlyons@email.arizona.edu)

1 Data Processing System Architecture Response CyVerse (UArizona) [1,2] currently operates an 8 petabyte (PB) iRODS [3,4] data grid distributed across 38 resource storage systems with replication of data at the Texas Advanced Computing Center (TACC) [5] that transparently provides rapid local access on TACC HPC systems and additionally serve as offsite backup. CyVerse connects over Internet2 [6] to the National Science Foundation (NSF) Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) [7] super computer framework of HPC, OpenScienceGrid HTC [8–10], and Jetstream-2 cloud resources [11,12]. CyVerse also connects to commercial clouds via Internet2 and Sun Corridor networks. CyVerse' iRODS data store has been in operation for over ten years and hosting doubled in size between 2017 and 2021 and is expected to double again by 2023.

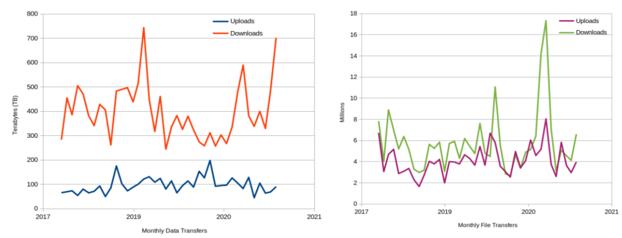


Fig 1a. Data Store usage (TiB/month) download and upload (left), number of files (n/month) (right).

CyVerse' users download (re-use) ~20 TiB per day and upload ~5 TiB new data per day via iRODS (Fig 1a). Specialized one-time transfers up to 50 TiB per day have been accomplished for specific projects using multiple iRODS connections with no interruptions in service to the rest of the platform. Public data downloads over HTTPS are supported via WebDAV caching [13–15]. Of specific relevance to this response are that multiple ontologies, metadata, and semantic query of data are managed via iRODS. Cloud buckets (S3/GCloud/Azure) alone do not provide the metadata attribution which are needed for faceted search or for reproducible research objects. iRODS resource servers create a **federated data grid,** providing the abstraction and unified and consistent path to data hosted with multiple commercial cloud providers or on-premise storage systems, allowing annotation and findability with structured queries and faceted search across CyVerse. The integration iRODS atop commercial cloud storage provides necessary structure to large data archives. Data processing is accomplished via a managed container orchestration environment consisting of HTCondor [8] and Kubernetes [16,17] that supports RESTful API services, such as TAPIS [18]. To support jobs to XSEDE and other resource providers in a data science workbench, called the Discovery Environment (DE) [19](Fig 1b) that manages interactive development environments (IDE) e.g., Jupyter, RStudio, Remote Desktops, and web based applications.

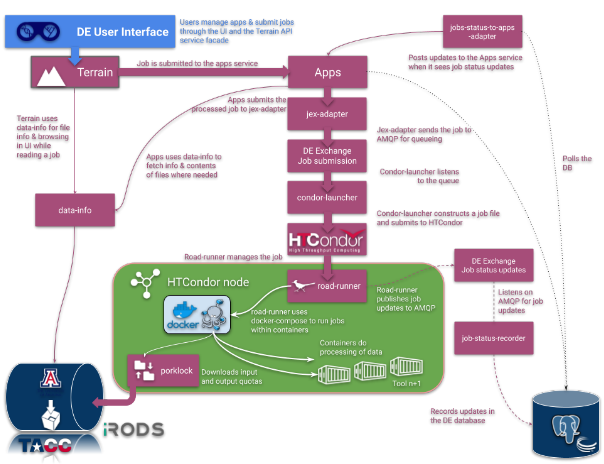


Fig 1b. Executable apps and workflows are managed by HTCondor and the CyVerse Terrain API [20,21]. Jobs are submitted through the DE interface where they trigger a job submission service managed by HTCondor with Advanced Message Queueing Protocol (AMQP) [22]. Once the job runs it is sent to a node where a program called road-runner uses Docker-Compose to manage the execution. Data are copied back to the iRODS data store when the app completes using a program called porklock. A PostgreSQL database monitors all job status and outcomes.

These data systems are access through CyVerse compute infrastructure, including the Discovery Environment and VICE, which are web-based platforms for users to integrate analytical tools, applications, and workflows. These systems support non-interactive/batch (e.g., traditional HPC/HTC applications) and interactive tools (e.g., Jupyter, RStudio). Tools are deployed using Docker containers and run across HTCondor and Kubernetes clusters (Fig 1b).

2 Open Science Response. Digital Research Objects, including data, analysis tools, and workflows [23,24] are managed through CyVerse using a combination of free and open-source software (FOSS) containers, container orchestration, workflow management systems, and event-based framework with continuous analysis (Fig 2a). CyVerse manages and simplifies the user experience of working on shared cyberinfrastructure. This simplicity helps reduce the digital divide with disadvantaged and underserved groups who do not have experience on cyberinfrastructure, can pay for access to commercial cloud computing, or provide their own technical support for managing their own cyberinfrastructure.

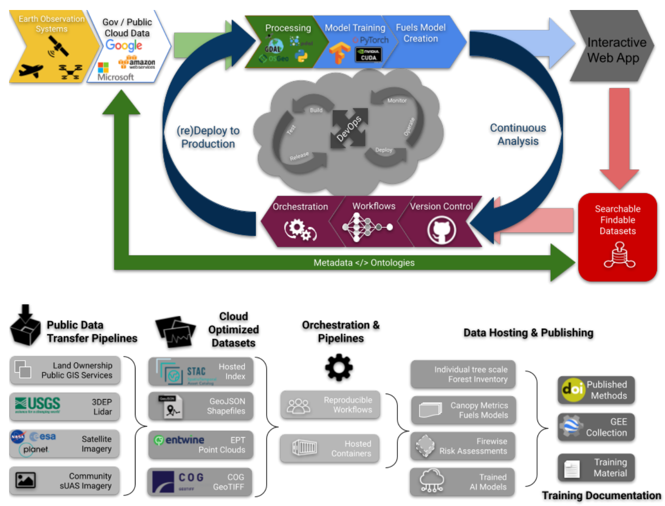


Fig 2a. Examples of remote sensing for mensuration of vegetation (fuels), wildland fire mitigation, and health monitoring [25] as a research object. Research object leverages public cloud-optimized and analysis-ready datasets [26] which are processed into derivative data products hosted on CyVerse DataCommons [27]. Methods are published using GitHub/Zenodo DOI, data with DataCite DOI on CyVerse DataCommons, and training material are made available online in version controlled markdown documentation.

Managing the entire research object for a specific research outcome requires that open science practitioners adopt reproducible methods from the onset until the end of their projects [23,24]. These methods include using version controlled container recipes with point-released versions of all software environments and metadata attribution of their creators and maintainers, containers hosted on public registries and their cryptographic hashes (SHA256) used for identification (Fig 2b).

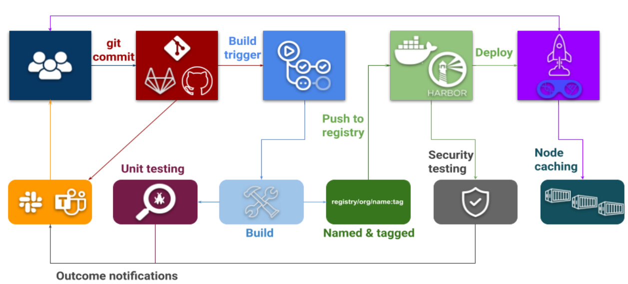


Fig 2b. Featured container deployment in CyVerse and CI/CD workflow. Image recipes (Dockerfiles) are hosted on GitHub/GitLab with DOI, and use build triggers with automation servers (GitHub Actions) to build and tag images. Tested images are pushed to public and private registries (DockerHub, Harbor). Images are cached on production servers (nodes) in CyVerse for rapid deployment as containers at runtime.

Analysis-ready cloud-native data [26] should be hosted publicly, without egress fees, and with accessible end points for researchers coming from institutional on-premises or public research major infrastructure (i.e. ACCESS) to do work on these data. Robust documentation for all parts of the research should be made available, these include: deployment of the cyberinfrastructure, development of the containerized software, launching and execution of the workflows, hosting of all data and algorithms used in the analyses (computational notebooks optional). Asynchronous training materials are a critical, and often overlooked component of reproducible research. Developing robust SOP and workshop materials help to increase adoption and drive community involvement. CyVerse partners with digital literacy groups to teach in-person, virtual, and asynchronous lessons for the bioinformatics and geoinformatics community. Participation with digital literacy groups like The Carpentries [28], Research Bazaar [29], PanGeo [29,30], and ESIP [31] are strongly advised.

3 Component Technologies Response. Vertically linking Earth Observation Systems data to local scale data (plot observations) and scale free data (genomics) requires a multi-use data science workbench which can manage computation and connect to large scale data storage. Onboarding novice cyberinfrastructure users who wish to bring their preferred scientific software (custom IDE, statistical software packages, algorithm stores, etc) into the shared environment requires a simplified user interface. On CyVerse this is accomplished through a managed containerization orchestration workspace (Fig 3a) which can be deployed anywhere, on either commercial cloud or public cyberinfrastructure resources.

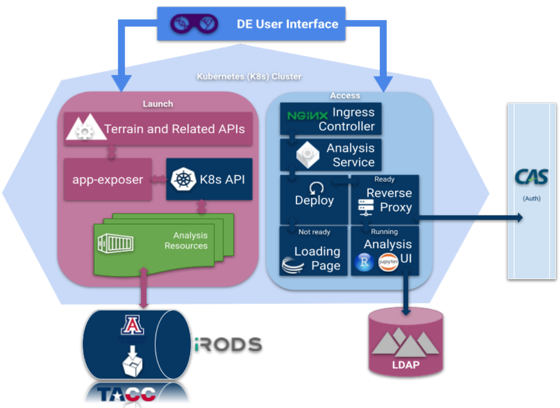


Fig 3a. The Discovery Environment (DE) manages interactive jobs through Kubernetes (K8s) and the Terrain API. Access to apps are managed by an Ingress Controller (NGINX), the analysis service shows whether the app is deployed, loading, or currently running and loads the UI for the analysis. Central authentication is managed by KeyCloak. Users can load data from the iRODS datastore into their running containers. LDAP manages the user's secure authentication. Data Store is actively mirrored at TACC from UArizona.

Fig 3b. CyVerse flexes between vertically scaling out to commercial cloud with large public data; to horizontal scaling with managed workflows for processing user contributed data, which vary from bioinformatics to geoinformatics on public research cloud, HPC and HTC.

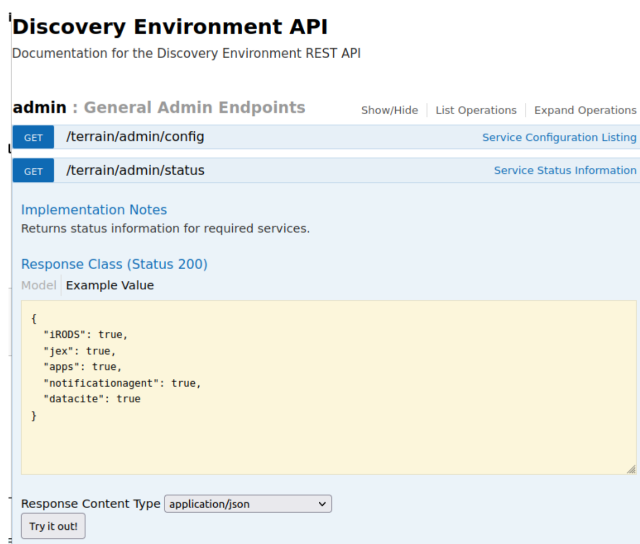
Data are often generated, collected, or deposited from multiple clouds or edge clouds, oftentimes requiring the ability to execute workflows near the originating cloud containing the data before forwarding primary or derivative products to another cloud. To enable multiple heterogeneous cloud compute, CyVerse Cacao [32] allows users to import, execute, and share multi-cloud workflows in the form of "infrastructure as code" (IaC) templates, such as Hashicorp Terraform [33], Argo Workflows [33,34], or Kubernetes using their own credentials. Multi-cloud workflows are stored and managed in public git repos, like GitHub.com or GitLab.com. Non-technical users can simply reuse existing curated workflows created by CyVerse staff or community-generated workflows stored within the CyVerse Cacao system. CyVerse Cacao users who use DataWatch can create "continuous analysis" [35] or automated workflows whenever their data changes in a continuous analysis pattern. CyVerse Cacao is one of the primary user interfaces for XSEDE Jetstream 2 [https://jetstream-cloud.org/], one of NSF's largest OpenStack multi-regional clouds and includes clouds located in Indiana University, TACC, Cornell University, Arizona State University and Hawaii. Private CyVerse infrastructure deployments on AWS GovCloud support projects such as the US Air Force Research Laboratories for space situational awareness (SSA) and near-earth object monitoring, and for Healthcare projects that manage and analyze protected patient data (HIPAA/PHI).

All components of CyVerse cyberinfrastructure have been demonstrated in practice for bioinformatics [36]; it has also recently facilitated breakthroughs for astronomical, earth and space sciences. Specifically, the Event Horizon Telescope (EHT) project hosts the results of its black hole modeling and imaging on CyVerse DataCommons [37–39] for which we grant the DataCite DOI and co-manage their results data on an external resource server. CyVerse hosts 400 TiB of model data for the Southern Ocean Carbon and Climate Observations and Modeling project (SOCCOM) [40] and UArizona Libraries geospatial archive.

4 Downstream Interoperability Response

Serving and analyzing data on commercial clouds in requester-pays buckets denies access to large communities of global researchers who are unfunded and uneducated in commercial cloud. This payee barrier only serves to increase the digital divide [41,42]. In particular, there is a common tale of unexpected or accidental cloud use and data transfer fees surprising novices who have not used these services before. Cloud-native analysis ready data formats that can use bit and range extractions reduce unnecessary data egress, i.e. CloudOptimizedGeotiff (COG), CloudOptimized PointClouds (COPC), HDR5, and GeoJSON selections have made significant impacts on reducing transfer sizes. Spatio-Temporal Asset Catalogs (STAC) API, and computational notebooks (RMarkdown, Jupyter) represent valuable new tools for the geospatial community and platforms like Microsoft Planetary Computer are in fact built around this new ecosystem of analysis ready data and notebooks. However, access to research-scale computing on these commercial cloud platforms are limited to underserved and historically disadvantaged communities. NSF's ACCESS and public cloud resources are available to help reduce the digital divide and to allow researchers from small and underserved institutions to work with data like those developed by NASA.

CyVerse facilities Interoperability across resources by developing, using (internally), and publishing its Science API, Terrain: https://cyverse-de.github.io/api/. This API is REST-based and used across CyVerse's main science workbench platform, the Discovery Environment, and all backend services. This API is open to all CyVerse users to aid in the development of novel CI resources and the interoperability across established CI. To help developers and researchers leverage Terrain, there are extensive documentation (linked above) and a web-based interface to describe, test, and create API calls: https://de.cyverse.org/terrain/docs/index.html.



NASA mission science data processing should focus on research object coproduction and open APIs with the requisite free and open source software in public version control, instructions for compilation across processors (x86, ARM64, etc), hosting of containers on public registries or privately managed registries with public access (e.g., GoHarbor.io)

Management of large corpuses of data on cloud requires they be given structure which makes them searchable and findable, and allow for users' defining inquiries, i.e., faceted search [43]. Existing frameworks for indexing [44,45], domain specific ontologies with Semantic Web query languages, e.g., SparQL, GeoSparQL [46]; ontology based metadata for remote sensing [47] linked to the Semantic Web. CyVerse has successfully used iRODS to annotate and provide searchable metadata through its DataStore.

5 Other Recommendations Response

The major players in geospatial for the research community are ESRI (ArcGIS), Google EarthEngine, and now Microsoft Planetary Computer. These licensed or free-tiered cloud services are valuable and provide easy to use IDE and GUI for the largest segment of the GIS community. However, vendor-lock in and changes to end user license agreements and free access policies create uncertainty to their long term benefit to the community. In addition, CyVerse has identified operationalized AI/ML, continuous analysis/model delivery, and continuous integration and being essential future needs for cyberinfrastructure enabled research discovery.

CHAPTER 3 WHAT IS MLOPS?

Continuous Model Delivery

This setup contains **pipelines** for automatic training of the deployed model as well as for speeding up the experimental process. Refer to Figure 3-3 for a graphical representation of this setup.

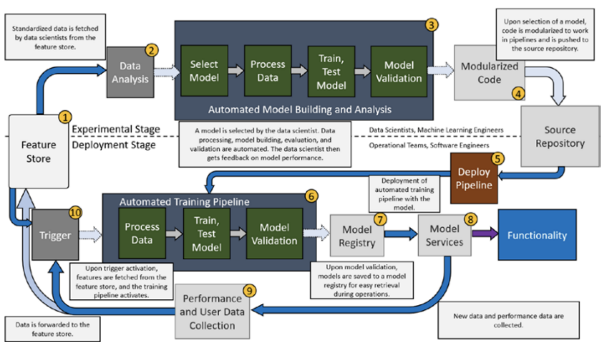


Figure 3-3. Graph depicting a possible deployment setup of a machine learning model with automation via pipelines

Continuous Integration/Continuous Delivery of Pipelines

In this setup, we will be introducing a system to thoroughly test pipeline components before they are packaged and ready to deploy. This will ensure **continuous integration of pipeline code** along with **continuous delivery of pipelines**, crucial elements of the automation process that the previous setup was missing. Refer to Figure 3-4 for a graphical representation of such a setup.

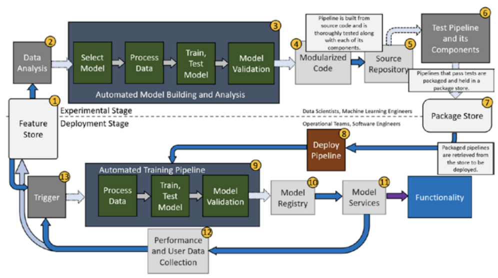


Figure 3-4. Graph depicting added testing systems and a package store to the automation setup in Figure 3-2

Figures taken from: https://link.springer.com/content/pdf/10.1007%2F978-1-4842-6549-9.pdf

To facilitate the processing of large datasets, CyVerse provides an event-driven service that allows users to monitor files or folders streaming into the CyVerse Data Store, called DataWatch [48]. Whenever a file or folder is modified, added, or deleted, DataWatch will perform one or more notification actions containing the list of change events. These user-configurable actions include posting to a webhook, writing a file to WebDAV or AWS S3, or emailing the list of changes to an address. The capability to send a list of data changes to a webhook provides an extensible "smart" data mechanism, by executing event-driven tools within CyVerse or downstream services to analyze data. CyVerse supports dozens of unique users every day and thousands a month, as it grows in user number and data volume we expect to reach PB-scale for internal networks and internet upload/download. In addition to DataWatch, CyVerse's is developing CACAO, https://gitlab.com/cyverse/cacao/-/blob/master/README.md, which provides the backend infrastructure for continuous analyses using Kubernetes Clusters. CACAO is accessed through RESTful microserve endpoints, ensuring flexibility, extensibility, and interoperability with other systems and CI resources.

1. Merchant N, Lyons E, Goff S, Vaughn M, Ware D, Micklos D, et al. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. PLoS Biol. 2016;14: e1002342.
2. Home. [cited 29 Jan 2022]. Available: https://cyverse.org
3. Hedges M, Hasan A, Blanke T. Management and preservation of research data with iRODS. Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience. New York, NY, USA: Association for Computing Machinery; 2007. pp. 17–22.
4. Xu H, Russell T, Coposky J, Rajasekar A, Moore R, de Torcy A, et al. iRODS Primer 2: Integrated Rule-Oriented Data System. Synthesis Lectures on Information Concepts, Retrieval, and Services. 2017;9: 1–131.
5. HOME - Texas Advanced Computing Center. [cited 29 Jan 2022]. Available: https://www.tacc.utexas.edu/
6. Home. In: Internet2 [Internet]. 5 Dec 2019 [cited 29 Jan 2022]. Available: https://internet2.edu/
7. Advanced Computing Systems & Services: Adapting to the Rapid Evolution of Science and Engineering Research. In: Beta site for NSF - National Science Foundation [Internet]. [cited 29 Jan 2022]. Available: https://beta.nsf.gov/funding/opportunities/advanced-computing-systems-services-adapting-rapid-evolution-science-and
8. Bockelman B, Cartwright T, Frey J, Fajardo EM, Lin B, Selmeci M, et al. Commissioning the HTCondor-CE for the Open Science Grid. J Phys Conf Ser. 2015;664: 062003.
9. Altunay M, Avery P, Blackburn K, Bockelman B, Ernst M, Fraser D, et al. A Science Driven Production Cyberinfrastructure—the Open Science Grid. Int J Grid Util Comput. 2011;9: 201–218.
10. Juve G, Rynge M, Deelman E, Vockler J-S, Berriman GB. Comparing FutureGrid, Amazon EC2, and Open Science Grid for Scientific Workflows. Comput Sci Eng. 2013;15: 20–29.
11. Hancock DY, Fischer J, Lowe JM, Snapp-Childs W, Pierce M, Marru S, et al. Jetstream2: Accelerating cloud computing via Jetstream. Practice and Experience in Advanced Research Computing. New York, NY, USA: Association for Computing Machinery; 2021. pp. 1–8.
12. Hancock DY, Stewart CA, Vaughn M, Fischer J, Lowe JM, Turner G, et al. Jetstream-Early operations performance, adoption, and impacts: Early Jetstream Performance and Results. Concurr Comput. 2018;57: e4683.
13. Surhone LM, Timpledon MT, Marseken SF. WebDAV: WebDAV, Hypertext Transfer Protocol, World Wide Web, Internet Engineering Task Force, Server (computing), Working Group, Operating System. Betascript Publishing; 2010.
14. squid : Optimising Web Delivery. [cited 29 Jan 2022]. Available: http://www.squid-cache.org/
15. Varnish HTTP Cache — Varnish HTTP Cache. [cited 29 Jan 2022]. Available: https://varnish-cache.org
16. Orzechowski M, Balis B, Pawlik K, Pawlik M, Malawski M. Transparent Deployment of Scientific Workflows across Clouds - Kubernetes Approach. 2018 IEEE/ACM International Conference on Utility and Cloud Computing Companion (UCC Companion). ieeexplore.ieee.org; 2018. pp. 9–10.
17. Bernstein D. Containers and Cloud: From LXC to Docker to Kubernetes. IEEE Cloud Computing. 2014;1: 81–84.
18. Projects. [cited 30 Jan 2022]. Available: https://tapis-project.org/projects/
19. Discovery Environment. [cited 30 Jan 2022]. Available: https://de.cyverse.org
20. cyverse-de. GitHub - cyverse-de/terrain: DE's main API entry-point service. In: GitHub [Internet]. [cited 30 Jan 2022]. Available: https://github.com/cyverse-de/terrain
21. cyverse. GitHub - cyverse/terrain-notebook: Jupyter Notebooks for interacting with the CyVerse DE API (terrain). In: GitHub [Internet]. [cited 30 Jan 2022]. Available: https://github.com/cyverse/terrain-notebook
22. Subramoni H, Marsh G, Narravula S, Lai P, Panda DK. Design and evaluation of benchmarks for financial applications using Advanced Message Queuing Protocol (AMQP) over InfiniBand. 2008 Workshop on High Performance Computational Finance. 2008. pp. 1–8.
23. Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I. Research Objects: Towards Exchange and Reuse of Digital Knowledge. Nature Precedings. 2010; 1–1.
24. Peng RD. Reproducible research in computational science. Science. 2011;334: 1226–1227.
25. Home. [cited 30 Jan 2022]. Available: https://promethean-gift.github.io
26. Abernathey RP, Augspurger T, Banihirwe A, Blackmon-Luca CC, Crone TJ, Gentemann CL, et al. Cloud-Native Repositories for Big Scientific Data. Computing in Science Engineering. 2021;23: 26–35.
27. Data Commons. [cited 30 Jan 2022]. Available: https://datacommons.cyverse.org
28. The Carpentries. In: The Carpentries [Internet]. [cited 31 Jan 2022]. Available: https://carpentries.org/index.html
29. Research Bazaar 2021. In: ResBaz Conference [Internet]. [cited 31 Jan 2022]. Available: http://resbaz.com//
30. Pangeo — Pangeo documentation. [cited 31 Jan 2022]. Available: https://pangeo.io/
31. ESIP. [cited 31 Jan 2022]. Available: https://www.esipfed.org/
32. CyVerse / CACAO. In: GitLab [Internet]. [cited 1 Feb 2022]. Available: https://gitlab.com/cyverse/cacao
33. Terraform by HashiCorp. In: Terraform by HashiCorp [Internet]. [cited 1 Feb 2022]. Available: https://www.terraform.io/
34. Argo Workflows - The workflow engine for Kubernetes. [cited 1 Feb 2022]. Available: https://argoproj.github.io/argo-workflows/
35. greenelab. GitHub - greenelab/continuous_analysis: Computational reproducibility using Continuous Integration to produce verifiable end-to-end runs of scientific analysis. In: GitHub [Internet]. [cited 1 Feb 2022]. Available: https://github.com/greenelab/continuous_analysis
36. ACOG Publications: December 2021. Obstet Gynecol. 2021;138: 949.
37. The Event Horizon Telescope Collaboration. First M87 EHT results: Calibrated data. CyVerse Data Commons; 2019. doi:10.25739/G85N-F134
38. Data Commons. [cited 26 Jan 2022]. Available: http://doi.org/10.25739/kejs-2n22
39. The EHT MWL Science Working Group. Multi-wavelength observations of M87 during the 2017 Event Horizon Telescope campaign. CyVerse Data Commons; 2021. doi:10.25739/MHH2-CW46
40. SOCCOM. [cited 29 Jan 2022]. Available: https://southernocean.arizona.edu/content/soccom
41. Jackson LA, Zhao Y, Kolenic A 3rd, Fitzgerald HE, Harold R, Von Eye A. Race, gender, and information technology use: the new digital divide. Cyberpsychol Behav. 2008;11: 437–442.
42. Norris P. The digital divide. Routledge; 2020.
43. Vandic D, van Dam J-W, Frasincar F. Faceted product search powered by the Semantic Web. Decis Support Syst. 2012;53: 425–437.
44. Welcome to Apache Solr. [cited 31 Jan 2022]. Available: https://solr.apache.org/index.html
45. OpenSearch. In: OpenSearch [Internet]. [cited 31 Jan 2022]. Available: https://opensearch.org/
46. Aldana-Martín JF, García-Nieto J, del Mar Roldán-García M, Aldana-Montes JF. Semantic modelling of Earth Observation remote sensing. Expert Syst Appl. 2022;187: 115838.
47. Fan J, Yan J, Ma Y, Wang L. Big data integration in remote sensing across a distributed metadata-based spatial infrastructure. Remote Sens (Basel). 2017;10: 7.
48. CyVerse / DataWatch. In: GitLab [Internet]. [cited 29 Jan 2022]. Available: https://gitlab.com/cyverse/datawatch