Issues in Data Sharing for Environmental Health Sciences

Mike Conway
NIEHS/NIH
mike.conway@nih.gov

Deep Patel
NIEHS/NIH
deep.patel@nih.gov

ABSTRACT

NIEHS is, like many other research enterprises, entering a new era of opportunity and challenge as methods of data-driven discovery emerge. The challenges of managing FAIR (Findable, Accessible, Interoperable and Reusable) (Hagstrom, 2014) data have multiplied as the variety, velocity and lifecycle requirements of environmental health data increase. FAIR data sharing relies on careful curation and management of upstream metadata quality, placement in appropriate data sharing environments and an expansive view of the 'data repository' as a constellation of general and specific data repositories that still exhibits the qualities of FAIR. The evolution of FAIR presents new opportunities and requires new ways of looking at the role of iRODS and policy-based data management.

Keywords

FAIR; data sharing; repositories; iRODS

INTRODUCTION

The National Environmental Health Association (NEHA) defines environmental health science as "the science and practice of preventing human injury and illness and promoting well-being" [1]. The National Institute of Environmental Health Sciences has the mission to "discover how the environment affects people in order to promote healthier lives." The importance of data and metadata management in service of this mission cannot be overstated. The NIEHS experience is not unique. Data management challenges, and the conceptual framework of policy-based data management to address these challenges have evolved over the years from data preservation in a single repository, through federated data sharing environments, to a new reality that is a network of heterogeneous generalist and specific repositories. The new distributed world supports FAIR data discovery and sharing, access controls, and distributed analysis.

IRODS AND EVOLVING DATA MANAGEMENT CHALLENGES

iRODS in the digital preservation era

If we trace the evolution of Policy-Based Data Management pioneered by the DICE Center, we see roots in the archival and data preservation community. iRODS evolved from the original Storage Resource Broker (SRB) and was redesigned as an open source data management platform with funding by the National Archives of the United States. In the original conceptual framework, the focus was on the application of management policies over the data life cycle in the form of a data grid, where a rule engine could enforce management policies in order to establish a trustworthy preservation environment. The data grid created a logical global namespace over a distributed storage environment. This logical view and the policy management focus made iRODS a compelling platform for data preservation, creating layers of abstraction and infrastructure independence from the underlying storage technology. From these capabilities Dr. Reagan Moore was able to form a “Theory of Digital Preservation”. This theory says that there are [2]:

- A minimal set of preservation processes (microservices),
- A minimal set of preservation metadata that can describe the preservation environment and effect of the preservation processes,
- A way of assessing the preservation metadata to validate that a repository is consistent (assessment criteria).

The theory posits that given these capabilities, an assertion can be made that a preservation environment is trustworthy. This model is a very concise framework for understanding the purpose of iRODS, even as the context has evolved. This model is still compelling in the world of FAIR data, where trustworthiness is a core requirement.

iRODS in federated research environments

The next era of iRODS, which we can all the “federation era”, was ushered in with the launch of the National Science Foundation’s DataNet initiative. The DataNet Federation Consortium was one of several DataNet initiatives funded by the National Science Foundation. DFC focused on the iRODS platform as the foundation for data federation. The DataNet Federation Consortium (DFC) retained the archival/repository view of the data grid and added multi-disciplinary sharing of scientific research data via federation mechanisms as a transformative element. DataNet identified functional areas that should be addressed in cyberinfrastructure for multi-disciplinary data-driven science. These areas were [3]:

1. data deposition, acquisition and ingestion,
2. metadata/ontology management,
3. data security,
4. data integration and interoperability,
5. data analysis and visualization.

DFC built on the notion of preservation capabilities and management policies from the data repository era and extended these concepts in a federation as means to automate many of the necessary tasks to serve the DataNet functional areas. DFC demonstrated that capabilities like indexing, publication, and data processing could be included in the scope of policy based data management [4].

In the DFC federation, multiple zones under different administrative control could be linked together, data could be shared between collaborators, and management policies could be enforced at each storage location. The policy managed framework in the context of a federated environment was extended further as a means to automate scientific workflows, including the segmentation of workflow operations such that these operations could be distributed among different parts of the federation. Moore and Rajasekar did acknowledge that more loosely coupled workflows would also be required, and interestingly proposed the capability to characterize the operations of a workflow, including the data management policies that were activated via a vocabulary as an aid to reproducibility. The authors made a prescient observation that the federation’s policy capabilities, including the possibility to stage a task at an appropriate storage resource, could work in tandem with loosely coupled workflows, saying “this style of integration of processing pipelines with collection-based data management is expected to become the basis for data-driven research”.

iRODS in disaggregated environments

We observe that iRODS is moving into a third era, and recent experiences at NIEHS reinforces this notion. We view this through the lens of our own data management challenges in Environmental Health Science. As we moved from the data preservation era to the federated data collaboration era, we saw that the "Theory of Digital Preservation" and the policy-based data management approach remained at the heart of the matter and this continues to be the case. The policy-based data management conceptual frameworks still apply and can help keep iRODS relevant.

There are several forces that characterize what we will call the "disaggregated environments" era. These are:

- The breaking of tightly coupled federation, replaced with multi-platform standards,
- The emphasis on highly structured data models,
- The focus on data associated with publications, with attendant focus on reproducibility and provenance,
- Cloud architectures and distributed data pipelines,
- The rise of containers and standard workflow languages,
- The requirement to manage sensitive data and honor data usage agreements.

The breaking of tightly coupled federation

As outlined above, iRODS has its roots in digital preservation and archiving. A data grid was a zone of control that provided a global logical namespace over multiple storage resources. In the federation era the focus shifted to federated zones representing multiple institutions, where iRODS still was the ultimate technology stack and policy and automation were largely under iRODS control. Federation remains a powerful mechanism for data sharing across collaborating institutions, however the data management challenges now extend far beyond the federation boundary.

In the NIH Strategic Plan for Data Science [5], much attention is paid to what is called the "biomedical research data-resource ecosystem". The strategic plan notes that there is a proliferation of repositories, including institutional repositories external to NIH, along with data repositories run by the various branches of NIH. In addition, there are multiple specialized repositories where subsets of research data are deposited. For example, a study may produce gene expression data, which is to be deposited into GEO, the Gene Expression Omnibus [6], while also producing sequencing data to be submitted into the Sequence Read Archive [7]. Meanwhile, data sets associated with a publication may end up deposited in a general repository, resulting in a patchwork of repositories associated with a study, publication, or research question. A great illustration of the problem is the existence of a guide to the diverse ecosystem of generalist and specific repositories relevant to biomedical research at NIH produced by the National Library of Medicine [8], currently populated with 66 entries for domain-specific repositories and a half-dozen generalist repositories. It is evident that federation models explored in the DFC era will not apply in such a fragmented ecosystem.

In response to this, iRODS should refine the policy-based data management philosophy, elevating policy management in two areas. First, the expansion of support for policy management of metadata is required, making metadata management a first-class concern. In this new era, especially as there is a greater focus on data harmonization and the employment of standard vocabularies and common data elements, policy-based management approaches can strengthen iRODS ability to serve as the canonical data and metadata preservation and management environment. Management policies can help maintain and validate metadata standards and automate many of the ingest and transformation tasks that underlie distribution of data to a network of loosely-coupled repositories. Second, the focus on policies as a means to automate tasks like indexing and publishing make iRODS an attractive solution for creating data coordination and data submission nodes that manage distribution while maintaining metadata that can link distributed data together.

The policy approach is well suited to the requirements of a canonical repository of data and associated metadata with a focus on metadata acquisition and quality control. In the NIEHS case this emphasizes the use of standard controlled vocabularies for curation and validation. Upstream data collection from sample and project submission along with the ingestion of data from laboratory information management (LIMS) systems and analysis workflows provides much-needed data curation. Policy-based metadata management enables researchers to generate valid submissions to various general and specific repositories, and NIEHS is currently providing tools for publishing to GEO (Gene Expression Omnibus). If iRODS can play a role in mediating data submissions, it can then retain accession information and pointers to various repositories where data is deposited, enabling FAIR data discovery across disaggregated storage locations. The COPO project [9] describes a data submission brokering service with a similar design goal. COPO provides general metadata curation tools, applying community developed vocabularies and ontologies for curation and validation and then uses these metadata to broker data submissions to target repositories. Instrumenting these publishing and distribution tasks allows foldback of additional metadata curation as well as allows establishment of links between source datasets and publications and analysis data sets in external repositories. This can aid in later cross-repository searching.

The emphasis on highly structured data models

The NIH has placed an emphasis on Common Data Elements (CDEs) as an important enabler for FAIR data sharing in a highly distributed data ecosystem. Rubenstein and McInnes [10] observed that:

> *One of the main obstacles to advancing biomedical research is the inability to exchange and share data and knowledge. This is the result of: data collected using different terminologies, databases being established with lack of interoperability and with no linkage between them, negative results and lessons learned not being shared, and resources (including funding and patient population) being used in duplicated efforts with no coordination and collaboration.*

To this end, NIEHS has convened an Environmental Language Health Collaborative to bring together interested parties to "advance community development and application of a harmonized language for describing Environmental Health Science (EHS) research" [11]. The development of CDEs and community developed ontologies and controlled vocabularies improves interoperability.

While the curation and validation of metadata using vocabularies and ontologies is critical, the data model of the repository itself and the relations between objects is equally important. For example, the Gen3 Data Commons explicitly build a commons architecture on highly structured data model, represented as a graph, and with services for indexing, submitting data, and searching the graph representation for data objects of interest [12, p. 3]. One of the base assumptions in Gen3 is that there are multiple commons connected by a core set of API and services. This is described as a "narrow middle" architecture [13]. These services include indexing and search capabilities and these capabilities are explicitly built around a structured data model. In the view of Grossman, this deep metadata model is one of the differentiating factors between a "data commons" and a "data lake", where a data lake is characterized as "when data are stored simply with digital IDs and metadata (shallow indexing), but without a data model" [14]. This data lake versus data commons distinction is important to weigh as metadata capabilities, such as metadata templates are being considered by the iRODS Consortium [15]. While the initial discussions are on flat schema for metadata, it is a matter to consider whether iRODS would need to support graph-like structures, or whether the platform would rather delegate this level of sophistication to an external service.

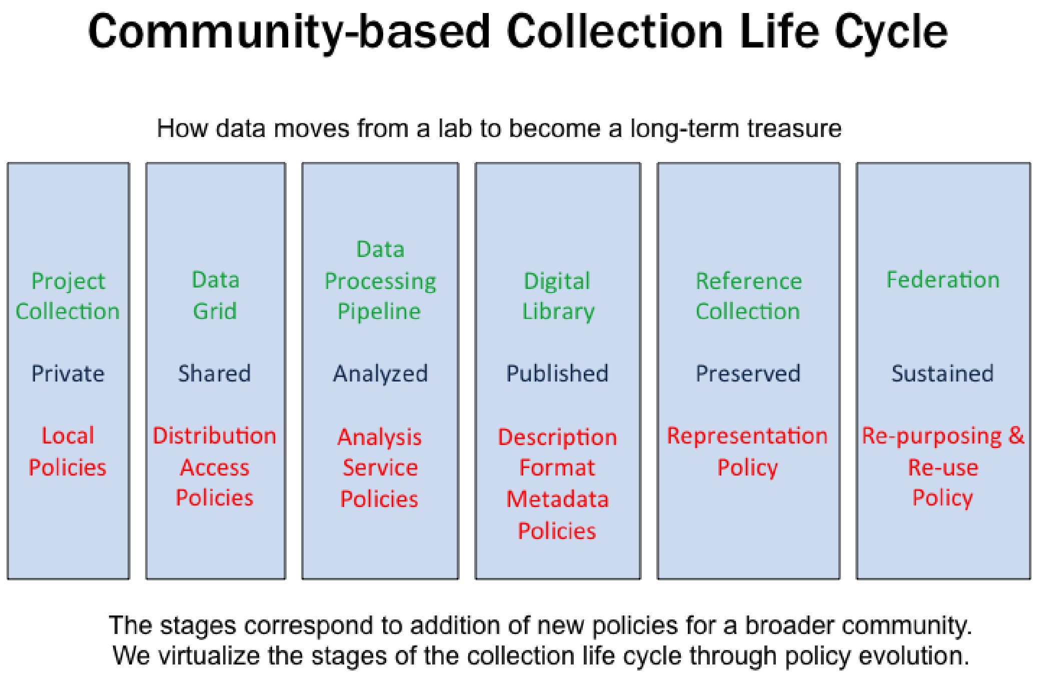
The focus on data associated with publications, with attendant focus on reproducibility and provenance.

Research data at NIEHS is utilized in many ways, and diverse research outputs are produced and published, ranging from data sets to reports to papers submitted to peer-reviewed journals. In all disciplines, there is a definite movement towards making research data a first-class concern on par with a published paper. When a conclusion or figure is presented in a scientific publication, the ability to cite the original data and to show how a conclusion or figure can be reproduced is now a requirement. Force11 has published a Joint Declaration of Data Citation Principals that captures this new reality [16], observing that "data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse." Viewed through the policy lens, these Force11 data citation principles fit well in the data lifecycle framework described by Moore and Rajasekar, where they observed that each stage of the research data lifecycle was governed by an explicit set of policies that represented a "community consensus on data sharing" [17]. The notion that iRODS can provide automation and policy control for management of research data sets at the publication side of the data lifecycle can in turn expand the sphere of control of management policies into the distribution and publication functions, expanding the scope of metadata and allowing better determination of the authenticity and reproducibility of research results, all the way back to the original data sets.

There are several challenges in the reproducibility aspects of FAIR that NIEHS is experiencing. Our data generation strategy is currently very fragmented. With the gradual formalization of data management at the Lab Information Management System (LIMS) level there are opportunities to extract metadata related to the various assays and procedures that were carried out. The Office of Data Science at NIEHS is championing efforts to formalize the vocabularies used to describe samples and assays and data handling in order to improve the upstream metadata quality. Data analysis (mapping to and updating the "Data Processing Pipeline" stage in the data lifecycle) is an area of difficulty, as the workflows and procedures used to analyze data are often detached from the control of any machine-actionable policies. There are three strategies we are focused on:

- Championing the use of notebooks as a way of capturing analysis,
- Championing the use of standard workflow languages and containerization to capture pipelines in detail,
- Considering aids and tooling for data submission for publication and to repositories as a way to capture links between publications and the source data.

Each of these strategies do not automatically establish the "R" in fair, however, they do present opportunities to improve the daily experience of analysts and researchers, improve the ability to share and record analysis steps for day-to-day work, and finally to instrument the analysis process for improved metadata and provenance capture. For example, Samuel et al. found that employing Jupyter notebooks enhanced the provenance capture for machine learning based analysis [18]. As we will note in our discussion of the rise of containers and standard workflow languages, the enhancements to portability, the ability to share or utilize best-practices pipelines, as well as the provenance and reproducibility benefits are significant. We saw in the COPO model how the publishing phase presents opportunities for management policies to gather provenance information. Expanding the ability of management policies to "see" activities and to gather metadata about processes and provenance can lead to better reproducibility of research results.



Community-based Collection Life Cycle

How data moves from a lab to become a long-term treasure

Project Collection	Data Grid	Data Processing Pipeline	Digital Library	Reference Collection	Federation
Private	Shared	Analyzed	Published	Preserved	Sustained
Local Policies	Distribution Access Policies	Analysis Service Policies	Description Format Metadata Policies	Representation Policy	Re-purposing & Re-use Policy

The stages correspond to addition of new policies for a broader community.
We virtualize the stages of the collection life cycle through policy evolution.

Figure 1. The research data lifecycle and relevant policies, from [17]

Cloud architectures and distributed data pipelines

One of the biggest disrupters to data management has been the rise of the cloud. While there are compelling capabilities in iRODS to connect to and manage cloud storage, there are also new challenges and opportunities in extending iRODS as a part of the broader architecture beyond storage. The notion of federation is present in discussions about cloud-based analysis but federation in these disaggregated environments is quite different, more along the lines of the “narrow middle” architecture with a sparse set of services that link otherwise independent environments. NIH, in their strategic planning, has identified several focus areas, including establishing researcher identity, managing data usage agreements, running computational tasks across multiple endpoints while enforcing fine grained authentication/authorization across multiple cloud providers as areas of concern. Each of these are natural areas where policy-based data management concepts can play a significant role.

It is the ability to rapidly scale to run complex analysis, taking advantage of the compute power and available tools, that is driving the migration of biomedical science to the cloud. The National Institutes of Health (NIH) acknowledged this migration in its original data commons pilot. The original pilot identified many challenges, including the portability of analysis to multiple cloud architectures, and the importance of minimizing ingress and egress charges, running an analysis task as close to the data as possible [19]. The Global Alliance for Genomics and Health (GA4GH) has worked to standardize workflows and pipelines, and to achieve the goals of portability and efficiency through the use of standard workflow languages, along with cloud-neutral data access and task execution standards [20]. Using standards such as the Data Repository Service (DRS) and Workflow Execution Service (WES), researchers have been able to demonstrate qualities of a federated data ecosystem [21].

The iRODS Consortium has considered “data-to-compute” and “compute-to-data” as general platform capabilities. Platforms such as CyVerse Discovery Environment have made great strides in providing “bring-your-own-compute” to their researcher community [22]. Strategies for iRODS to enable of these narrow-middle services and support emerging standards and conventions to support distributed analysis are a rich area for exploration. There has been prior work using iRODS to support distributed analysis on cloud platforms under the umbrella of the original NIH Data Commons Pilot as well as under the sciDAS program [23] which demonstrated the ability to distribute tasks based on a calculation of performance and cost factors. This work parallels many of qualities of the current GA4GH federated analysis projects, including the Task Execution Service (TES) [24]. These examples show that there may be multiple ways that iRODS could serve as a platform for distributed data pipelines.

The rise of containers and standard workflow languages

In large part, the ability to run pipelines and analysis in distributed environments relies on the use of containers such as Docker or Singularity. The CyVerse Discovery Environment is a great example within the iRODS Community, where researchers can create and containerize tools, annotate them, and add them into the research environment to run and share. Repositories such as BioContainers [25] and Dockstore [26] allow researchers to share containerized tools and workflows written in standard workflow languages such as CWL [27] and NextFlow [28]. The portability and reproducibility of containers is well understood. Formalizations for running containerized tasks on data at remote storage locations (as CyVerse does to some extent) using standard TES-compliant interfaces may be an interesting and beneficial capability.

The requirement to manage sensitive data and honor data usage agreements

If iRODS and policy-based data management have any intersection with capabilities that are desired in the proposed NIH Data Commons architecture, they would intersect with the need for fine-grained and dynamic decisions on data access. Efforts are underway to develop ontologies that can describe data usage , as well as standard systems, built on existing standards such as OpenID Connect and OAuth, for identifying researchers. There are efforts at NIH and at GA4GH to create something akin to a data passport. This system was described by Cabili et al [29] as a “Library Card”:

> *A Library Card would encode identity and other attributes of a researcher as a set of standardized claims. For example, institutional affiliation may be recorded in a claim, along with a level of assurance regarding the claim: (a) self-assertion (b) institutional email (c) proof of support of the claim by an authorized third party within the institution. These claims, and others like them, are recorded by the Library Card issuer and provided over secure protocols to relying parties who will in turn make access decisions based upon these claims*

iRODS policies and the ability to make decisions based on role, status, or dynamic factors such as acknowledgment of terms or data usage agreements would be a natural area to investigate, especially were iRODS is mediating access to research data sets. NIEHS did demonstrate a Data Repository Service (DRS) implementation that can run on native iRODS [30, p.]. This DRS implementation does have, in later planned iterations, the ability to provide GA4GH Data Passports [31] that can provide a framework for exploring how iRODS policies can interact with GA4GH Data Passports.

CONCLUSIONS

This paper observes that a third "era" of policy-based data management has arrived, driven by the disaggregated nature of repositories, the rise of the cloud, and the need to create federations across disparate systems for data-driven science. The capabilities demonstrated and developed in prior eras of iRODS of have not disappeared. However, the challenge and opportunity presents itself to build upon the success of the policy-based data management model and the core capabilities in the iRODS platform and learn how they apply and can advance the platform in this new federation era.

REFERENCES

[1] Global Alliance for Genomics and Health, "Definitions of Environmental Health | National Environmental Health Association: NEHA." https://www.neha.org/about-neha/definitions-environmental-health (accessed Jul. 12, 2021).

[2] R. Moore, "Towards a theory of digital preservation," Int. J. Digit. Curation, vol. 3, no. 1, pp. 63–75, 2008.

[3] J. W. Lee, J. Zhang, A. S. Zimmerman, and A. Lucia, "DataNet: An emerging cyberinfrastructure for sharing, reusing and preserving digital data for scientific discovery and learning," AIChE J., vol. 55, no. 11, pp. 2757–2764, Nov. 2009, doi: 10.1002/aic.12085.

[4] R. W. Moore and A. Rajasekar, "Reproducible Research within the DataNet Federation Consortium," p. 8.

[5] "NIH Strategic Plan for Data Science."

[6] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," Nucleic Acids Res., vol. 30, no. 1, pp. 207–210, Jan. 2002, doi: 10.1093/nar/30.1.207.

[7] R. Leinonen, H. Sugawara, M. Shumway, and on behalf of the International Nucleotide Sequence Database Collaboration, "The Sequence Read Archive," Nucleic Acids Res., vol. 39, no. suppl_1, pp. D19–D21, Jan. 2011, doi: 10.1093/nar/gkq1019.

[8] "Data Sharing Resources." https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html (accessed Jul. 13, 2021).

[9] A. Etuk et al., "COPO: a metadata platform for brokering FAIR data in the life sciences," bioRxiv, 2019, doi: 10.1101/782771.

[10] Y. R. Rubinstein and P. McInnes, "NIH/NCATS/GRDR® Common Data Elements: A leading force for standardized data collection," Contemp. Clin. Trials, vol. 42, pp. 78–80, May 2015, doi: 10.1016/j.cct.2015.03.003.

[11] "Environmental Health Language Collaborative - Harmonizing Data. Connecting Knowledge. Improving Health.," National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/research/programs/ehlc/index.cfm (accessed Jul. 13, 2021).

[12] "Gen3 - Set up Gen3." http://gen3.org/resources/operator/ (accessed Jul. 13, 2021).

[13] R. L. Grossman, "Progress Towards Cancer Data Ecosystems," Cancer J. Sudbury Mass, vol. 24, no. 3, pp. 122–126, 2018, doi: 10.1097/PPO.0000000000000318.

[14] R. L. Grossman, "Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data," Trends Genet., vol. 35, no. 3, pp. 223–234, Mar. 2019, doi: 10.1016/j.tig.2018.12.006.

[15] The iRODS Consortium, TRiRODS: Metadata Templates in iRODS. Accessed: Jul. 13, 2021. [Online Video]. Available: https://www.youtube.com/watch?v=_b4AvhhG7mc&ab_channel=TheiRODSConsortium

[16] M. Crosas, “Joint Declaration of Data Citation Principles - FINAL,” FORCE11, Oct. 30, 2013. https://www.force11.org/datacitationprinciples (accessed Jul. 14, 2021).

[17] R. W. Moore, A. Rajasekar, M. Conway, W. Schroeder, and M. Wan, “White Paper: National Data Infrastructure for Earth System Science,” White Pap. Submitt. EarthCube Proj. Httpearthcube Ning Comgrouptechnology-Resolut. Last Access 24 May 2013, 2011, Accessed: May 29, 2015. [Online]. Available: http://semanticommunity.info/@api/deki/files/13791/=003_Moore.pdf

[18] S. Samuel, F. Löffler, and B. König-Ries, “Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles,” ArXiv200612117 Cs Stat, Jun. 2020, Accessed: Jul. 14, 2021. [Online]. Available: http://arxiv.org/abs/2006.12117

[19] National Institutes of Health, “NIH Data Commons Pilot Phase.” Jun. 16, 2017. [Online]. Available: https://commonfund.nih.gov/sites/default/files/rm-17-026_commonspilotphase.pdf

[20] “GA4GH Cloud Workstream.” https://www.ga4gh.org/work_stream/cloud/#proposed-solution (accessed Jul. 14, 2021).

[21] “Realising the Genomics Ecosystem: ELIXIR Plays Key Role in GA4GH 2020 Connection Demos,” ELIXIR, Sep. 30, 2020. https://elixir-europe.org/news/realising-genomics-ecosystem-elixir-plays-key-role-ga4gh-2020-connection-demos (accessed Jul. 15, 2021).

[22] U. K. Devisetty, K. Kennedy, P. Sarando, N. Merchant, and E. Lyons, “Bringing your tools to CyVerse Discovery Environment using Docker,” F1000Research, vol. 5, p. 1442, Jun. 2016, doi: 10.12688/f1000research.8935.1.

[23] F. Jiang, C. Castillo, and S. Ahalt, “A Cloud-Agnostic Framework for Geo-Distributed Data-Intensive Applications,” p. 10.

[24] “GA4GH TES API: Bringing compatibility to task execution across HPC Systems, the Cloud and Beyond.” https://www.ga4gh.org/news/ga4gh-tes-api-bringing-compatibility-to-task-execution-across-hpc-systems-the-cloud-and-beyond/ (accessed Jul. 15, 2021).

[25] F. da Veiga Leprevost et al., “BioContainers: an open-source and community-driven framework for software standardization,” Bioinformatics, vol. 33, no. 16, pp. 2580–2582, Aug. 2017, doi: 10.1093/bioinformatics/btx192.

[26] B. D. O’Connor et al., “The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows,” F1000Research, vol. 6, p. 52, Jan. 2017, doi: 10.12688/f1000research.10137.1.

[27] P. Amstutz et al., “Common Workflow Language, v1.0,” Jul. 2016, doi: 10.6084/m9.figshare.3115156.v2.

[28] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, “Nextflow enables reproducible computational workflows,” Nat. Biotechnol., vol. 35, no. 4, pp. 316–319, Apr. 2017, doi: 10.1038/nbt.3820.

[29] M. N. Cabili et al., “Simplifying research access to genomics and health data with Library Cards,” Sci. Data, vol. 5, no. 1, p. 180039, Mar. 2018, doi: 10.1038/sdata.2018.39.

[30] “michael-conway/irods-ga4gh-dos: GA4GH Data Object Service for iRODS,” GitHub. https://github.com/michael-conway/irods-ga4gh-dos (accessed Jul. 21, 2021).

[31] “linking DRS with Passport Visa · Issue #339 · ga4gh/data-repository-service-schemas,” GitHub. https://github.com/ga4gh/data-repository-service-schemas/issues/339 (accessed Jul. 21, 2021).