Scaling HPC Education

Susan Mehringer*
Center for Advanced Computing,
Cornell University
shm7@cornell.edu

Mary P. Thomas*
San Diego Supercomputer Center,
University of California San Diego
mpthomas@ucsd.edu

Kate Cahill
Ohio Supercomputer Center
khill42@gmail.com

Charlie Dey
Texas Advanced Computing Center
charlie@tacc.utexas.edu

David Joiner
Kean University
djoiner@kean.edu

Richard Knepper
Center for Advanced Computing,
Cornell University
rich.knepper@cornell.edu

John-Paul Navarro
University of Chicago, Argonne
National Lab
navarro@anl.gov

Jeaime H. Powell
Texas Advanced Computing Center
jpowell@tacc.utexas.edu

ABSTRACT

Throughout the cyberinfrastructure community there are a large range of resources available to train faculty and young scholars about successful utilization of computational resources for research. The challenge that the community faces is that training materials abound, but they can be difficult to find, and often have little information about the quality or relevance of offerings. Building on existing software technology, we propose to build a way for the community to better share and find training and education materials through a federated training repository. In this scenario, organizations and authors retain physical and legal ownership of their materials by sharing only catalog information, organizations can refine local portals to use the best and most appropriate materials from both local and remote sources, and learners can take advantage of materials that are reviewed and described more clearly. In this paper, we introduce the HPC ED pilot project, a federated training repository that is designed to allow resource providers, campus portals, schools, and other institutions to both incorporate training from multiple sources into their own familiar interfaces and to publish their local training materials.

KEYWORDS

education, training, community engagement, survey

1 INTRODUCTION

We introduce the HPC ED pilot project, a federated training repository of vetted and tested training that is designed to allow resource providers, campus portals, schools, and other institutions to both incorporate training from multiple sources into their own familiar interfaces and to publish their local training materials. As a needs assessment prior to proposing HPC ED, the project team conducted a survey in October-November 2022 of providers of HPC training materials and related resources. The results [6, 18] showed that most respondents were both interested in, and able to, work toward community efforts to share and discover materials (see Section 2.) The HPC ED federated repository has been designed to identify and vet training resources from this broad set of offerings and to provide metadata and characterization that support successful discovery and utilization of training resources, and their incorporation into portals for research computing groups at universities, HPC centers, schools, domain-centered institutions and elsewhere. HPC ED also provides an API so that local centers will be able to include content identified in the repository and offer it to their local communities side-by-side with local offerings. In addition, local centers can upload training offerings that have been vetted and share with the broader computational science community.

HPC ED extends and enhances the ability of universities, departments, research computing groups, HPC centers, and domain-specific collaborations to discover and incorporate relevant and proven training materials into their own websites, portals, and science gateways, and to contribute to the federated repository. Leveraging of the federated training repository allows communities access to advanced CI-related training materials without the burden of creating and maintaining large sets of materials, and facilitate the professional development of individuals served by those institutions. In addition to providing sets of training materials that are commonly used for local activities, the federated repository will facilitate discovery of new materials that add to the overall catalog.

In this paper we describe our efforts to identify the needs of the HPC training community via our surveys (Section 2) which provides motivation for the HPC ED project. Section 3 presents an overview of the architecture of the Federated Training Catalog, our approach to developing the metadata, and an API for publishing and discovering the catalog content, and our commitment to maintaining community driven standards of quality. Section 4 describes our efforts to date to train our community to utilize the catalog by hosting workshops and hackathon and working with early adopters,

*Both authors contributed equally.

https://doi.org/10.22369/issn.2153-4136/15/1/8

who will provide important feedback for the Federated Training Catalog system. The paper concludes with a discussion that by our plans to help build a community of HPC trainers, we will help to ensure the longevity of the materials and metadata used for the catalog (Section 5).

2 SURVEY RESULTS OVERVIEW

The survey [6, 18] conducted in 2022 was conducted to learn if there was a benefit to improving how HPC training materials are shared and discovered. The survey showed that while community members are successful at finding materials, there are many barriers that make it difficult; the most cited reason was difficulty in finding materials at the right depth or level, as shown in Table **??**.

Table 1: What Barriers Have You Encountered When Searching for Materials?

Barriers encountered	Responses
I can't find materials on the topic I need	35
I can find materials on the topic, but not at the depth or level I need	72
I find too many materials, and I can't effectively sort through them all	44
I am aware of specific appropriate materials, but search engines don't list them in the top results	26
Other	28
Total	93

In Figure 1, we explore whether respondents are both interested in making data discovery easier, and able to provide metadata. We see in the lower right quadrant that very few respondents want to make finding data easier, but lack the ability. The greatest number of respondents, in the upper right quadrant, have both the interest and ability. This shows great potential in the community moving forward with solutions. Altogether, the results imply that the community sees the potential for improving discovery of materials and many have the interest and ability to contribute to a solution.

3 FEDERATED ARCHITECTURE

We are building an architecture designed to enable organized and collective HPC training material sharing and discovery across the national and international research and education community. This community currently relies on two approaches for HPC training material discovery. First, larger research projects and education-focused organizations often have local training catalogs and discovery portals where members of their community discover materials selected for relevance and value to their community. These catalogs may contain locally produced training material or manually selected and entered information about external training resources. Commercial organizations, such as LinkedIn Learning, also provide their own catalogs and training discovery environments. Second, internet search engines like Google and streaming services like YouTube are often used to discover HPC training materials.

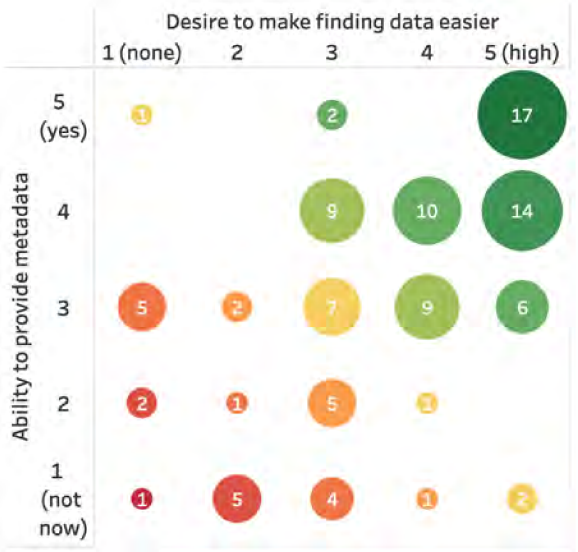

Figure 1: Results of two survey questions: Does your organization want to make it easier for the public to find your training and education materials? & Would your organization be willing and able to share your training and education materials in a public catalog by providing metadata about your materials in a standard format?

At the core HPC ED is a federated training catalog that (1) leverages and builds on the strengths and flexibility of organization-specific training catalogs and portals, (2) addresses the many deficiencies of search engines and streaming media services, (3) enables every individual looking for HPC education material, whether they are in an organization providing local training material or not, to discover training material across all federated training catalog participants,

HPC ED provides an API where resources can be published to the catalog and queries can be made to identify and retrieve content via the metadata system. Organizations that produce training materials and events will be able to publish into the Federated Catalog and reach a greater audience (for example MathWorks leveraging the catalog to publish documentation and events). Conversely, organizations that are curating appropriate materials for their local community can browse the catalog and add resources to their local portal. Organizations can both share their material and discover new materials to be added to local portals.

An overview of the federation process is presented in Figure 2. In this workflow, Training Developers/Curators use the API to submit a request to publish their materials and events into the Federated Catalog. Once approved, the information is entered into the catalog and made available to the public on the catalog site via the API query interface. This catalog of resources has the potential to encompass thousands of training resources and events that will be made available through an API that allows sites to add resources to their local portals and share resources with the catalog.

Using the HPC ED API, organizations seeking training are able to browse the catalog for materials and add them to their local

portals and information systems. In this way, they can present complete training material sets without the difficulty of creating and maintaining them consistently over time. By leveraging a federated model, the training community can highlight the best possible training resources and emphasize competencies developed individually.

The base technology for the repository has been tested and is complete, and is currently running behind the https://software.xsede.org and the https://research.illinois.edu sites, both of which provide information about HPC software and research resources respectively. The former includes training materials on a much broader basis (including LinkedinLearning.com resources). HPC ED provides additional quality assurance of resources and integration into HPC learning roadmaps. The project team will establish a similar repository for training materials that collects information about the materials: location information, title, and metadata about content and topics. A feature of this system is that training materials remain in their original location and are discovered via the repository itself or from within an HPC Center website that uses the HPC ED repository via an API.

3.1 Metadata, Taxonomy and Ontologies

Working with existing community efforts to build a set of standardized minimal HPC training metadata is critical for publishing and discovering training information effectively. The Research Computing and Data community is active in this area, but the current lack of consistent metadata for HPC training is a major barrier to effective discovery. A standard taxonomy of HPC/CI training concepts, developed by an HPC training community, would make materials more easily searchable and discoverable, more readily adoptable and it will support the FAIR principles (Findable, Accessible, Interoperable, and Reusable) [11] for sharing of training materials. The HPC ED federated catalog builds on a foundation of standardized minimal HPC training metadata for publishing and discovering training information and will merge and standardize the HPC learning metadata from among our partnering organizations into a common metadata schema. [22] [15] [19] [20] [21] [9]

We propose that our effort begin with two types of metadata for elements in the federated repository: first, metadata that describes the training material, its access methods, and educational characteristics, including Title, Description, Authors, Publisher, Type, Language, Cost, Format, License, Target Group, Expertise Level, Certification details, and very importantly, Persistent Identifiers, Tags, or Keywords; and second, metadata that identifies the publisher and source of the training material so that when an individual selects a specific training item, they can be directed to the source catalog that published that material to browse all available information and to access that training material. Additionally, we will start with the Research Data Alliance (RDA) “Recommendations for a minimal metadata set to aid harmonized discovery of learning resources” that addresses many of the use cases and needs around basic training sharing and discovery and supports FAIR practices" [12]. Note that once the community can agree on this metadata, we can begin exploring or connecting to other HPC based taxonomies or ontologies efforts.

A key outcome of this project is the formation of and participation in an HPC/CI training materials ontology community (e.g., an NSF ACCESS Affinity Group [2] [4]). This is described in Section 5 We are in the process of forming a Metadata Team of collaborators who will iteratively review community schema standards and identify discovery terms that need to be added to the schema, and post periodic schema upgrades for public use. This activity will be most likely organized through the ACCESS MATCH Affinity program.[4] The "HPC Ontologies and Metadata" Affinity group was created in Sept, 2023.[3]

Defining, categorizing, and standardizing the metadata is a significant effort. There are currently several efforts in the HPC/CI area that are building ontologies that describe the HPC ecosystem, but there is no single/primary metadata set, taxonomy, or ontology [7] [8] [14] [17] [23] [16] [24]. This is because of the complexities of the hardware, software, other components, organizations, local admin policies, etc. Where possible, we will identify and use existing, common metadata sets, taxonomies, and ontologies. Where needed, we will identify and add new terms to these existing ontologies and work with existing communities to update them or to develop the HPC/CI training materials ontology.

3.2 Sharing/Publishing Materials

To enable individuals, organizations, and projects with local training material to share and publish, HPC ED provides a secure publishing RESTful API. Using this API, any authorized organization will be able to automatically publish standard metadata from their local training materials catalog to the Federated Catalog. Affiliated organizations will implement automated publishing once, and rerun/synchronize frequently to refresh published information about their local training materials. Published training metadata will be stored in the Federated Catalog where it is aggregated with training metadata from other organizations. We will track who published each training element, perform basic quality checks, and inform the publishers of metadata quality issues.

We foresee challenges to the long-term maintenance of the project, such as how to handle material that is no longer being updated or is no longer accessible. We plan to address these challenges using quality assurance (see Section 3.4), and also by tracking when metadata was published and most recently refreshed. Working with an advisory board and the HPC community, the HPC ED project team will conduct periodic quality reviews of content to assess for relevance and correctness of materials. Publishers will be required to automatically refresh their metadata, and if they fail to do so within a configurable and reasonable interval, their metadata will be expired in the Federated Catalog. To support projects, organizations, and researchers that have training materials but do not have local training catalogs, we plan to coordinate with the ACCESS Support (e.g. MATCH) project [5] which already offers a way to publish training materials into the ACCESS reference material catalog.

3.3 Discovery of Materials

To enable organizations, projects, and individuals to discover published training material we will implement a federated training discovery/search RESTful API. This API will provide advanced

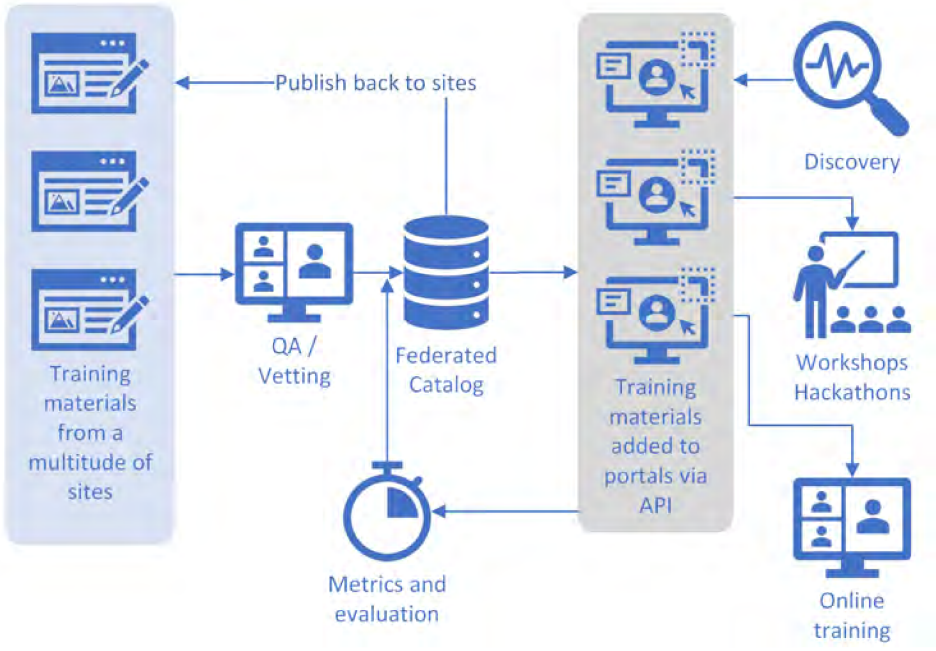

Figure 2: Training Developers/Curators use the API to publish their materials and events into the Federated Catalog. Organizations seeking training use the API to browse the catalog for materials and add them to their local portals.

search capabilities. For example, it will enable individuals to perform precise searches on specific metadata values, find materials from a particular organization, author, or targeted skill level, as well as perform more advanced key terms-based searches that rely on our related HPC training material ontology and taxonomy work. Advanced searching will ensure that the most relevant results are returned for search terms and enable relevance rankings based on known relationships between terms.

Projects and organizations with their own catalogs and training discovery portals or websites will be able to present their users with training materials published and shared by other organizations through the Federated Catalog. They will be able to do this by building into their websites the ability to query the catalog using our APIs and present their users with federated training search results.

To support projects, organizations, and researchers that do not have local training portals we hope to partner with the ACCESS Support (MATCH) project to implement a way for those individuals to discover training through a MATCH community-wide portal.

3.4 Quality Assurance

For the HPC training community, building a repository is not enough; we must also assure the accuracy of metadata of items shared through the federation. This includes verifying the status and nature of materials, validating their accuracy, and accrediting that metadata associated with the materials is appropriate [13]. Some of these processes can be fully automated, others assisted by artificial intelligence techniques, and some are simply human labor. QA will include processes that use automation where appropriate and build on the crowd-sourced nature of input from users of the repository and at workshops and hackathons through rating information from the community on the material in the catalog including a 5-star rating system, and monitoring the existence and uptime of links.

An additional component of the review process will center on material metadata. While review metadata that is collected will at the simplest level implement a star rating and user comments, additional feedback collected will focus on whether the metadata being displayed accurately reflects the catalog item. This can include descriptive information, such as author, title, and source, as well as audience level and content description. By sharing back this review information with catalog maintainers, we will provide a value add for adopters of the federation. Catalog maintainers will be able to opt-in to the review system.

4 PROJECT TRAINING

4.1 Workshops and Hackathons

The 2022-2026 National Science Foundation Strategic Plan [10] notes that development of human capital must begin with training that embeds generations of technical expertise followed by cultural/ community capital. For the HPC ED project, that means leveraging workshops in a phased manner to engage with MSIs, non-research, and academic research institutions, industry vendors, and research organizations to ingest training resource data needed to fill and maintain the catalog. Additionally, hackathons (time-scoped deliverable-driven events), and later Birds of a Feather

(BoF) sessions at targeted community conferences (see Section 5 will engage the collective human capital to maintain, evaluate, and identify community technical training needs. To clarify, the use of the workshops, hackathons, and BoFs for the purpose of collecting, maintaining, and building both the resources connections and fostering community engagement with the Federated Catalog each will have targeted outcomes within the three training delivery phases. These workshops, hackathons, and BoFs will be held virtually and, when applicable, in person.

4.2 Early Adopters

Early adopters will be encouraged to participate at all levels of development of the Federated Training Repository system, and will provide valuable feedback for needed changes. This group will serve as alpha clients and will define and/or redefine methodologies, processes, and initial user interface templates based on their required experiences. The reason this will be targeted for a workshop is to allow the development team rapid turnaround from critical path concerns identified by the user group. A diverse set of early adopters is essential to ensure broader engagement, current knowledge resources, and post- proposal funding resource opportunities respectively ensuring the culture of the project is inclusive. The project team has received letters of commitment from a number of sites willing to be eager adopters and help contribute to the holdings of the repository. These projects include enhancing coursework at MSI institutions, training Cyberinfrastructure Professional (CIP), hosting training catalogs at our local institutions, and collaborations with various ACCESS projects.

5 BUILDING AND SUSTAINING COMMUNITY

A community-wide project can only thrive when it has input, feedback, and use by the community. We have organized avenues for community engagement and communication on the project activities, which include our training program described above, and holding meetings with key stakeholders. A working group has been formed within the ACM SIGHPC Education Chapter to discuss metadata standards for sharing materials across all interested organizations. A BoF was held at PEARC23 to gather community input of what is needed to make existing training materials to be more findable, accessible, interoperable and reusable (FAIR)[21] for the whole community to benefit from them. Two key outcomes resulted from this meeting: (1) the participants wanted to meet as often as possible, synchronizing with other meetings; and (2) an affinity group within ACCESS [3] was recently created as a result of the discussions at this BoF. Future events include meeting at SC23, and the Science Gateways 2023 Annual Conference, and other meetings. For more information, visit the HPC ED website, located at: https://github.com/HPC-ED/HPC-ED.io.

We have a number of partners who confirmed their interest and intention in integrating their training resources as early adopters, who are committed to helping us to grow the repository and to collect feedback on the product and procedures. We share regular updates through the HPC ED Google Group mailing list and newsletter to allow those who want to know about activities can be kept updated. To join the mailing list, send an email to hpc-ed@googlegroups.com.

ACKNOWLEDGEMENTS

Our thanks for collaborative efforts and events go to the SIGHPC Education Chapter [1]. We want to acknowledge the use of several NSF funded resources and services including: the SDSC Expanse project (#1928224); the TACC Stampede System (# 1663578); NSF CyberTraining: Pilot: HPC ED: Building a Federated Repository and Increasing Access through Cybertraining (# 2320977); the Extreme Science and Engineering Discovery Environment (XSEDE) (NSF award #ACI-1548562); and the NSF ACCESS Track 3 Award: COre National Ecosystem for CyberinfrasTructure (CONECT) (#2138307);

REFERENCES

[1] 2023. SIGHPC Education chapter. https://sighpceducation.acm.org/
[2] ACCESS. 2023. ACCESS Advanced Cyberinfrastructure Coordination Ecosystem. https://access-ci.org/
[3] ACCESS. 2023. ACCESS HPC Ontologies and Metadata" Affinity group. https://support.access-ci.org/affinity_groups
[4] ACCESS. 2023. ACCESS Support Affinity Groups. https://support.access-ci.org/affinity_groups
[5] ACCESS. 2023. ACCESS User Support. https://support.access-ci.org/
[6] K Cahill, D Joiner, S Lathrop, S Mehringer, and A & Navarro, J-P & Weeden. 2022. Final Results: National Survey on Educational and Training Materials Repositories. https://www.cac.cornell.edu/about/pubs/Survey2022.pdf
[7] Gabriel G. Castañé, Huanhuan Xiong, and Dapeng Dong & John P. Morrison. 2018. An ontology for heterogeneous resources management interoperability and HPC in the cloud. Future Generation Computer Systems 88 (2018), 373–384. https://doi.org/10.1016/j.future.2018.05.086
[8] Dong Dai, Yong Chen, Philip Carns, John Jenkins, Wei Zhang, and Robert Ross. 2019. Managing Rich Metadata in High-Performance Computing Systems Using a Graph Model. IEEE Transactions on Parallel and Distributed Systems 30, 7 (2019), 1613–1627. https://doi.org/10.1109/TPDS.2018.2887380
[9] M. Emani and X Liao, C. & Shen. 2021. HPCFAIR: An Infrastructure for FAIR AI and Scientific Datasets for HPC Applications. https://custom.cvent.com/DCBD4ADAAD004096B1E4AD96F3C8049E/files/event/1fe48ee7ca1949c0b6ebd5f4a81c3d5f/04c21e2a6378405586b3e7ce51570e0b.pdf
[10] National Science Foundation. 2022. U.S. National Science Foundation 2022-2026 Strategic Plan. https://www.nsf.gov/pubs/2022/nsf22068/nsf22068.pdf
[11] Leyla Garcia, Bérénice Batut, Melissa L. Burke, Mateusz Kuzak, Fotis Psomopoulos, Ricardo Arcila, Teresa K. Attwood, Niall Beard, Denise Carvalho-Silva, Alexandros C. Dimopoulos, Victoria Dominguez Del Angel, Michel Dumontier, Kim T. Gurwitz, Roland Krause, Peter McQuilton, Loredana Le Pera, Sarah L. Morgan, Päivi Rauste, Allegra Via, Pascal Kahlem, Gabriella Rustici, Celia W.G. Van Gelder, and Patricia M. Palagi. 2020. Ten simple rules for making training materials FAIR. PLoS Computational Biology 16, 5 (2020), 1–9. https://doi.org/10.1371/journal.pcbi.1007854
[12] N. J. Hoebelheinrich, K. Biernacka, M. Brazas, L. J. Castro, N. Fiore, M. Hellström, E. Lazzeri, E. Leenarts, P. M. Martinez Lavanchy, E. Newbold, A. Nurnberger, E. Plomp, L. Vaira, and A. van Gelder, C. W. G. & Whyte. 2022. Recommendations for a minimal metadata set to aid harmonised discovery of learning resources. https://doi.org/10.15497/RDA00073
[13] David Joiner, Steven Gordon, Scott Lathrop, Marilyn McClelland, and D. E. Stevenson. 2005. Applying Verification, Validation, and Accreditation Processes to Digital Libraries. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05). Association for Computing Machinery, New York, NY, USA, 382. https://doi.org/10.1145/1065385.1065485
[14] R.M. Keller, D.C. Bemos, R.E. Carvalhol, D.R. Hall, S.J. Rich, I.B. Sturken, and S.R. Swanson, K.J. & Wolfe. 2004. SemanticOrganizer: A Customizable Semantic Repository for Distributed NASA Project Teams. , 5 pages. https://ntrs.nasa.gov/api/citations/20040084377/downloads/20040084377.pdf
[15] Richard M Keller, Daniel C Bemos, Robert E Carvalhol, David R Hall, Stephen J Rich, Ian B Sturken, Keith J Swanson, and Shawn R Wolfe. 2004. SemanticOrganizer : A Customizable Semantic Repository for Distributed NASA Project Teams. , 15 pages. https://doi.org/10.1007/978-3-540-30475-3_53
[16] C. Liao, P.-H. Lin, G. Verma, T. Vanderbruggen, M. Emani, and X. Nan, Z. & Shen. 2021. HPC Ontology: Towards a Unified Ontology for Managing Training Datasets and AI Models for High- Performance Computing. , 69-80 pages. https://www.osti.gov/servlets/purl/1832325
[17] L. Ma, J. Mei, Y. Pan, K. Kulkarni, and A. Fokoue, A. & Ranganathan. 2007. Semantic Web Technologies and Data Management. https://www.w3.org/2007/03/RdfRDB/papers/ma.pdf
[18] Susan Mehringer, Kate Cahill, Scott Lathrop, Charlie Dey, Mary Thomas, and Jeaime H Powell. 2023. Assessing Shared Material Usage in the High Performance Computing (HPC) Education and Training Community. The Journal of

Computational Science Education (2023).
[19] Shodor. 2023. HPC University Resources Page. Retrieved September 8, 2023 from http://hpcuniversity.org/resources/search/
[20] Diana Tanase, David A. Joiner, and Jonathan Stuart-Moore. 2006. Computational science educational reference desk: A digital library for students, educators, and scientists. D-Lib Magazine 12, 9 (2006), 0–4. https://doi.org/10.1045/september2006-tanase
[21] G Verma, M Emani, C Liao, P Lin, T Vanderbruggen, X Shen, and B Chapman. 2021. HPCFAIR: Enabling FAIR AI for HPC Applications. https://www.osti.gov/biblio/1830466
[22] Alexander Willner, Mary Giatili, Paola Grosso, Chrysa Papagianni, Mohamed Morsey, and Ilya Baldin. 2017. Using Semantic Web Technologies to Query and Manage Information within Federated Cyber-Infrastructures. Data 2, 3 (2017). https://doi.org/10.3390/data2030021
[23] L. Youseff and D. Butrico, M. & Da Silva. 2008. Toward a Unified Ontology. https://ieeexplore.ieee.org/document/4738443
[24] Aolong Zhou, Kaijun Ren, Xiaoyong Li, Wen Zhang, Xiaoli Ren, and Kefeng Deng. 2021. Semantic-based discovery method for high-performance computing resources in cyber-physical systems. Microprocessors and Microsystems 80 (2021), 103328. https://doi.org/10.1016/j.micpro.2020.103328