Building a Federated Catalog for CyberTraining Materials: The HPC-ED Pilot Project

Mary P. Thomas
mpthomas@ucsd.edu
University of California, San Diego
La Jolla, California, USA

Susan Mehringer
Center for Advanced Computing,
Cornell University
Ithaca, New York, USA
shm7@cornell.edu

Katharine Cahill
New Jersey Institute of Technology
Newark, NJ, USA
katharine.cahill@njit.edu

Charlie Dey
Texas Advanced Computing Center
Austin, TX, USA
charlie@tacc.utexas.edu

Brian Guilfoos
Ohio Supercomputer Center
Columbus, OH, USA
guilfoos@osc.edu

David Joiner
Kean University
Union, NJ, USA
djoiner@kean.edu

Richard Knepper
Center for Advanced Computing,
Cornell University
Ithaca, New York, USA
rich.knepper@cornell.edu

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

Jeaime H. Powell
Texas Advanced Computing Center
Austin, TX, USA
jpowell@tacc.utexas.edu

ABSTRACT

To improve the sharing and discovery of CyberTraining materials, the HPC-ED Pilot project team is building a platform for the community to better share and find training materials through a federated catalog. The platform, currently in early test mode, is focused on a flexible platform, informative metadata, and community participation. By creating a framework for identifying, sharing, and including content broadly, HPC-ED will: allow providers of training materials to reach new groups of learners; extend the breadth and depth of training materials; and enable local sites to add or extend local portals.

CCS CONCEPTS

• **Applied computing** → **Education**; **Digital libraries and archives**; *Document searching*; • **Information systems** → *Search engine architectures and scalability*.

KEYWORDS

Education, Training, Community engagement, HPC, Cyberinfrastructure, Metadata, Globus

ACM Reference Format:
Mary P. Thomas, Susan Mehringer, Katharine Cahill, Charlie Dey, Brian Guilfoos, David Joiner, Richard Knepper, John-Paul Navarro, and Jeaime H. Powell. 2024. Building a Federated Catalog for CyberTraining Materials: The HPC-ED Pilot Project. In *Practice and Experience in Advanced Research Computing (PEARC '24), July 21–25, 2024, Providence, RI, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3626203.3670586

PEARC '24, July 21–25, 2024, Providence, RI, USA

ACM ISBN 979-8-4007-0419-2/24/07
https://doi.org/10.1145/3626203.3670586

1 INTRODUCTION

The current landscape of cyberinfrastructure (CI) and high-performance computing (HPC) training materials is distributed across multiple organizations and portals, creating challenges for educators and learners to identify up-to-date and relevant training materials. The HPC-ED Pilot project team is building an international federated catalog to improve the sharing and discovery of CI and HPC materials, and to foster the formation of a community to better share and find these materials. [15, 16] The primary goal of the federated catalog is to significantly scale the ability of researchers, educators, and students to more effectively and efficiently find and use relevant training material.

Compounding the difficulty of finding appropriate materials are incomplete and inconsistent material descriptions. At the same time, those who create and maintain training materials want to make it easier to find and access their materials, and to reach new audiences to CI and HPC information. By creating a framework for identifying, sharing, and including content broadly, HPC-ED will: allow providers of training materials to reach new groups of learners; extend the breadth and depth of training materials; and enable local sites to add or extend local portals. Labeling materials with more accurate metadata, enabling training material owners to share their materials by publishing metadata to the catalog, and enabling institutions and projects to enhance their portals by adding training materials shared by others through the catalog will help achieve key project goals.

To achieve these aims, the HPC-ED project is building a collaborative framework and community for discovering and sharing CI and HPC training and education materials to ensure that high-quality materials are available throughout the community and can be easily integrated with local websites.

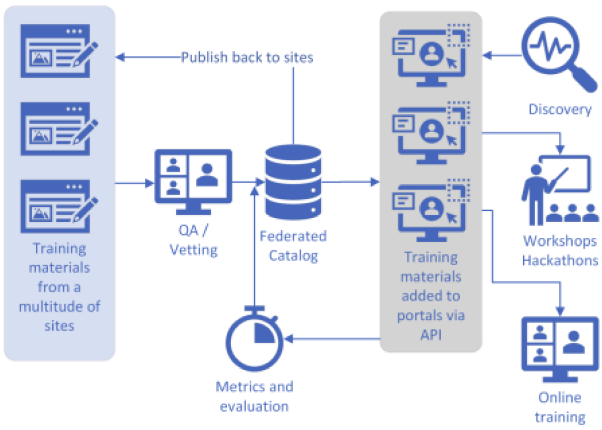

Figure 1: Overview of the proposed processes for the Federated Training Materials Catalog, including publishing, searching, and storage.

2 HPC-ED ARCHITECTURE AND PROCESS OVERVIEW

The HPC-ED project is currently in its pilot phase (as of September 2023) and we're working to develop the tools to make this federated catalog available to the HPC education & training community. Figure 1 describes the proposed process: providing a catalog for training providers to publish materials and researchers to discover them.

2.1 HPC-ED Metadata

Metadata is a computer science term, which has the following features: a "shorthand representation of the data to which it refers"; allows for the easy retrieval, management, and use of data; and is structured to model the most important features of the data it describes. [2] HPC-ED bases its metadata set on the Research Data Alliance (RDA) recommended minimal metadata set to aid in harmonized discovery of learning resources. [8] The fields used by the HPC-ED system for publishing and searching the catalog is detailed in the online documentation. See the HPC-ED website for updates and more detailed information on the metadata. [12]

The RDA recommended minimal metadata is by design the minimal useful metadata needed to enable "harmonized" discovery across multiple catalogs. It is not intended to be a comprehensive collection of metadata. The recommended minimal metadata specifically includes enough information for individuals searching for training material to find relevant training, and enables the person viewing the minimal metadata to either view the complete and comprehensive metadata *in* the catalog that the minimal metadata came from, or to directly access the training resources itself.

2.2 Tools for Searching and Publishing

HPC-ED uses multiple catalogs to support development and testing efforts. We anticipate a single production catalog shared by all collaborators. HPC-ED catalogs are public and can be searched without authorized credentials. To browse HPC-ED catalogs (Globus Indexes) and their metadata, clients can view the developer's portal (see Section 3.1.1. [17] Publishing requires authorized credentials. We are currently issuing a single publishing (writer) credential to each partner or beta testing institution.

2.2.1 Publishing Metadata. Publishing Metadata involves generating HPC-ED specified metadata in JSON (JavaScript Object Notation [14]) format and publishing it to an HPC-ED Globus Search index, described in detail on the project wiki page. [12] Note that the use of JSON is required by Globus. Globus documentation refers to this process, seen in Figure 2, as ingesting entries into an index. There is not enough space in this paper to describe the JSON files in detail, but that information can be found on the HPC-ED Wiki pages.[13]

2.2.2 Searching Metadata. Searching Metadata involves sending a query to the Globus Search service that should be run against an HPC-ED Catalog (a.k.a. Globus Index). Various parameters may be specified in the query to determine what filters should be applied to retrieve results, and which facets (or content) should also be returned to facilitate further narrowing of search results. Since HPC-ED catalogs are public, no authorization credentials are needed to search HPC-ED catalogs. As shown in Figure 2, metadata can be searched using the ACCESS Operations Search Pilot Portal, the *Globus CLI*, Python SDK, SDK for Javascript, and the *Globus API*.[7] Links to each of these methods can be found on the HPC-ED wiki pages. [12]

3 BUILDING A COMMUNITY

A critical component of the HPC-ED project is building community. To develop expansive community collaborations and coordination, we are conducting an outreach campaign to maximize community contributions, usage of training materials, resulting in more effective impact. Toward this effort, we are presenting papers and conducting BoFs at conferences, running hackathons, building a mail list, sending regular newsletters, arranging coordinated topics with ACM HPC EDU committees, and meeting with ACCESS regularly. [11]

Importantly, community participation in HPC-ED furthers training impact in several ways: increasing the number of resources in the federated catalog, increasing discovery and usage of materials, and sharing community-developed tools, to help others build and augment local portals. To further build community, we partnered with two ACCESS organizations: ACCESS Support enables the sharing and discovery of research computing training material; and ACCESS Operations has a platform designed for sharing and discovery of metadata rated to research computing.

Early adopters are key to a successful pilot project because they provide critical roles in several areas: development and testing of the software; define realistic and practical project goals, and can help build community. In addition to adding valuable resources to the catalog, they inform improvements to documentation and procedures, and some will produce tools to share. Early adopters are those who committed to helping the project at the proposal stage and others who learned of the project through our outreach activities and joined the mail list. Early adoption activities began in

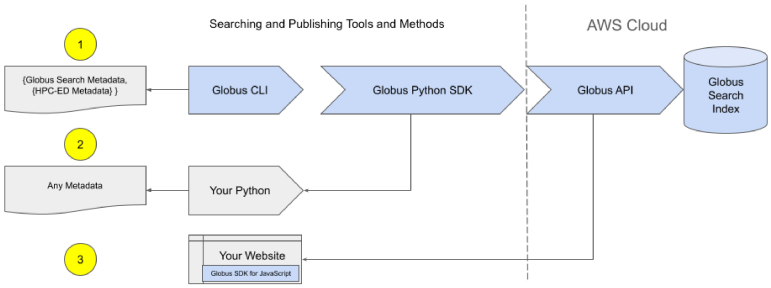

Figure 2: Diagram showing the tools or methods used to publish (steps 1-2) and to search and display HPC-ED metadata (steps 1-3).

Spring 2024 with an email announcement, providing documentation and two online sessions: a practical overview, [9] followed by an "office hours" session.

3.1 Implementation Examples

The projects described below serve as examples and testcases in learning what work needs to be done to work with the API described above in Section 2. It is also hoped that these projects can serve as how-to examples for future projects wishing to contribute materials, or extract materials for their training programs.

3.1.1 HPC-ED Developers Portal. To develop and test the HPC-ED search and publishing APIs, we developed a demo portal based on the django-globus-portal-framework (dgpf). The design of this portal is intended be very simple, and to serve as an internal project "test harness" as opposed to a full featured, fully navigable, user friendly portal (such as the ACCESS portal).[6] Reference information for HPC-ED *dgpf* customizations and configurations can be found in our GitHub repo. [19] The ability to select and search one of several catalogs is needed as we expect that during Alpha and Beta testing we will have several search catalogs with progressively higher-quality metadata. An example of a demo portal interfacing to one of the catalogs is shown in Figure 4 (left).

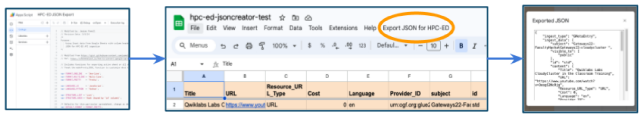

Figure 3: Google Sheet Extension Workflow for JSON Generation: Apps Script view (left), Google Sheets with "Export JSON for HPC-ED" extension(middle), and example JSON output window (right)

3.1.2 HPC Training Catalog. At the San Diego Supercomputer Center (SDSC), all training events and materials are hosted online using databased material. [20] Once an event is over, the event materials are moved to a training catalog and accessed through the Interactive Video web pages. [21] The web page includes simple metadata searching. As part of the pilot program, we have built a prototype portal based on ingesting existing and future events and extracting events from the catalog.

The process of ingesting existing training materials and presenting them in the demo portal is straightforward (see Figure 4, center). The basic steps included: (1) convert existing training data (which is stored in JSON files) to HPC-ED JSON format; (2) Upload the JSON files to the catalog using a Python script; (3) Search the catalog for selected materials and download the JSON files; (4) Update the portal software to read and display the selected HPC-ED JSON files. The ingestion of data is updated daily with a cron job. To date, we have ingested materials from around 80 events.

Future plans include: updating the production catalog to use HPC-ED formatted JSON files; pushing all new events to the HPC-ED catalog; expanding the search capabilities in the portal; and extracting training data from other organizations and displaying this on the portal.

3.1.3 Ingesting Material from Existing Training Material Sources. The Cornell Virtual Workshop (CVW) hosts asynchronous training materials via an online portal. [4] It contains about 40 online tutorials comprised of about 200 topics. Each topic includes a JSON file containing metadata. To facilitate ingesting the material into the catalog, we wrote a C# program that pulls the relevant information from each JSON file and writes out a single multi-resource input JSON file in the prescribed format. That JSON file is then used to share the material metadata by using the documented Globus commands to write the metadata to the federated catalog, using the Globus ingest command. The detailed procedure is shown below. Future plans include improving our resource metadata, sharing additional materials, and discovering gap material to include in our local portal.

(1) Read the HPC-ED Publishing documentation, paying special attention to metadata fields. [12]
(2) Augment and clean up JSON files containing topic metadata.
(3) Test sharing one resource: Create JSON file, check file for proper ID tags and syntax errors, e.g. with a simple tool like https://jsonlint.com, then ingest the file using the HPC-ED Publishing documentation. After uploading, check the developer's portal. [17]
(4) Share all resources: Write a C# program to create a multi-resource input file, then ingest and test as in the previous step.

3.1.4 Contributing Material using JavaScript and Google Sheets Extension. This project was developed to test contributing resources to the HPC-ED developers portal from the perspective of a resource publisher with only a few materials to add. Additionally, the inclusion of a Google Form for data ingestion into a Google Sheet was considered as a possible workflow. With that scope in mind, the creation of a JavaScript-based Google Sheet Extension was

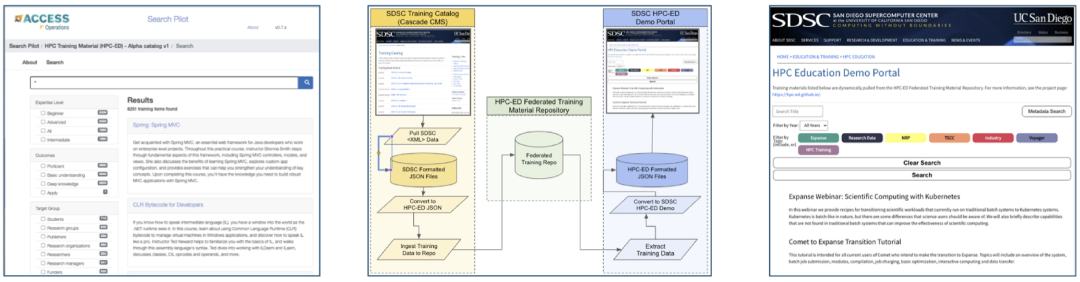

Figure 4: Images show the HPC-ED development portal (left), SDSC demo portal architecture and workflow (center), and the portal using HPC-ED catalog material (right).

developed and tested using resources from the Texas Advanced Computing Center of the University of Texas at Austin (TACC).[18] This method required the installation of the *Globus CLI* and terminal use to complete the authentication and use of the *Globus API for JSON* ingestion of the generated metadata.[7] The development process included the creation of a Google Sheet with column headers that matched the required data fields as defined by the HPC-ED "Publishing Metadata" and "Metadata Description" Wiki instructions [12]. The Google Sheet extension was written in the "Apps Script" manager in Google Sheets. JavaScript code was modified to generate the appropriate HPC-ED Globus JSON format. [5] Results generated JSON text as seen in Figure 3. The output was then ingested into the metadata catalog for publication using the Globus CLI.

3.1.5 Development of Scripts to Extract Materials from a Research Publication Resource . This project serves as a testcase in learning what work needs to be done to ingest existing materials from an existing publication website. The Journal of Computational Science Education JOCSE is an online publication featuring articles on the use of modeling and simulation and other computational science tools in the classroom. It currently houses 177 articles over 15 volumes. JOCSE is maintained as a Jekyll site, hosted on GitHub. Each article is stored under a directory structure that includes volume and issue, and metadata for each article is housed as header information in YAML format. We wrote a python script to descend the directory structure and parse each file, store each metadata element in a dictionary object, rename fields if they did not already align with the Globus required metadata and HPC-ED recommended metadata entries, correct for variations in date-time format, and perform other formatting and special character remove as needed. Globus subjects were assigned based on volume, issue, and article number. This was then written to a JSON file, and ingested using the Globus CLI.

Some issues we found in this test case included decisions on using optional metadata fields and suggested vocabularies that did not fit JOCSE metadata. For example, some metadata on the education level collected with JOCSE articles did not fit into the vocabulary for expertise level in the HPC-ED metadata vocabulary, requiring a choice between omitting some optional metadata, or not using the recommended vocabulary. Ultimately 14 of the metadata fields were used for ingest.

4 CONCLUSIONS AND FUTURE WORK

The HPC-ED team has implemented the federated catalog and associated metadata schema, supporting the publication and sharing of materials by the HPC-ED user community. HPC-ED is an NSF CyberTraining Pilot Project (#2320977). The API has been designed to be straightforward to use. Several large HPC centers are testing and implementing the API and data, and others have expressed interest in using it as evidenced by people who have joined our Google group [10] or responded to our survey. [3]

The HPC-ED community is growing as a result of several activities: workshop presentations and publications (SEHET23 [15], BPHET23Scaling-HPC-Education-2023, PEARC24 [22]; training tutorials; and holding BoFs at relevant meetings (PEARC23, PEARC24); a GoogleGroup with over 50 members; working with HPC/CI training providers; developing an ACCESS affinity group. In addition, we have developed the GitHub website with information and links to a training wiki.[13]

Moving forward we will improve the usability of the catalog interface, bring new partners who provide content, and explore ways to provide consistent listings that are easily integrated with partners' websites. As part of our future plans, we will be presenting a PEARC24 tutorial, titled "Publishing and Discovering Cybertraining Materials Across the HPC and CI Research Communities". The review function of the federated catalog will collect information on the quality of materials that will support highlighting relevance and accuracy. As the catalog matures, we will add tools that check for availability of materials and ADA and usability compliance and explore methods to implement LLM's based on the catalog metadata and supporting information.

ACKNOWLEDGMENTS

Our thanks for collaborative efforts go to the SIGHPC Education Chapter [1]. This project is funded through the NSF CyberTraining: Pilot: HPC ED: Building a Federated Repository and Increasing Access through CyberTraining (#2320977); and supported by multiple NSF grants (#2230127, #2017767, #2320934, #2112606). We appreciate being granted access to NSF resources and services including: the NSF ACCESS Track 3 Award: COre National Ecosystem for CyberinfrasTructure (CONECT) (#2138307); and several HPC resources (SDSC Expanse #1928224 and TACC Stampede # 1663578);

REFERENCES

[1] ACM. 2023. *SIGHPC Education chapter.* https://sighpceducation.acm.org/

[2] Encyclopedia Britannica. 2024. Advanced Computing Training Program. https://www.britannica.com/.

[3] 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

[4] Center for Advanced Computing. 2024. Cornell Virtual Workshop. https://cvw.cac.cornell.edu.

[5] Pamela Fox. 2024. *GitHub Gist - export.js.* https://gist.github.com/pamelafox/1878143/

[6] Globus. 2024. Django Globus Portal Framework (dgpf). https://github.com/globus/django-globus-portal-framework.

[7] Globus. 2024. Globus Search Service. https://docs.globus.org/api/search/.

[8] 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

[9] HPC-ED. 2024. *Early Tester Training.* https://www.youtube.com/watch?v=fD9g5z5qZFQ

[10] HPC-ED. 2024. *Google Group.* https://groups.google.com/g/hpc-ed

[11] HPC-ED. 2024. *HPC-ED Events.* https://hpc-ed.github.io/events

[12] HPC-ED. 2024. *HPC-ED Wiki Documentation.* https://hpc-ed.github.io/wiki

[13] HPC-ED. 2024. *Wiki.*

[14] JSON. 2024. *JavaScript Object Notation.* https://www.json.org/

[15] 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. In *Sixth Workshop on Strategies for Enhancing HPC Education and Training (SEHET23).* https://doi.org/10.22369/issn.2153-4136/14/2/4

[16] Susan Mehringer, Mary P Thomas, Charlie Dey, Kate Cahill, David Joiner, Richard Knepper, and Jeaime H Powell. 2023. Scaling HPC Education. In *Tenth SC Workshop on Best Practices for HPC Training and Education BPHTE23.* Denver, 41–46. https://doi.org/10.22369/issn.2153-4136/15/1/8

[17] ACCESS Operations. 2024. HPC-ED Developers Portal. https://search-pilot.operations.access-ci.org/.

[18] Jeaime Powell. 2024. *GitHub Gist - sheets-to-hpc-ed-JSON.* https://gist.github.com/jeaimehp/2656730f3cd021b59aa845e6f8d483e9

[19] HPC-ED Pilot Project. 2024. HPC-ED Customized DGPF Portal Repository. https://github.com/HPC-ED/hpc-ed_django-globus-portal-framework.

[20] SDSC. 2024. Advanced Computing Training Program. https://www.sdsc.edu/education_and_training/training_hpc.html.

[21] SDSC. 2024. Interactive Video Tutorials. https://education.sdsc.edu/training/interactive.

[22] Mary P Thomas, Susan Mehringer, Katharine Cahill, Charlie Dey, Brian Guilfoos, David Joiner, John-paul Navarro, Jeaime H Powell, and Richard Knepper. 2024. Building a Federated Catalog for CyberTraining Materials : The HPC-ED Pilot Project. *Accepted - Proceedings of PEARC 24* 1, 1 (2024), 1–6.