A Documentation Catalog for the AI Ecosystem: Design of the ICICLE AI Training Catalog

Amit Vyas
Department of Computer Science, The Ohio State University
vyas.154@osu.edu

Carlos Guzman
Department of Computer Science, The Ohio State University
guzman.109@osu.edu

Neelima Savardekar
Department of Computer Science, The Ohio State University
savardekar.2@osu.edu

Mary Thomas
San Diego Supercomputer Center (SDSC), UC San Diego
mpthomas@ucsd.edu

Michelle Miller
University of Wisconsin–Madison
mmmille6@wisc.edu

Alfonso Morales
University of Wisconsin–Madison
morales1@wisc.edu

Joe Stubbs
Texas Advanced Computing Center (TACC), University of Texas at Austin
jstubbs@tacc.utexas.edu

Christian Garcia
Texas Advanced Computing Center (TACC), University of Texas at Austin
cgarcia@tacc.utexas.edu

Beth Plale
University of Oregon
plale@uoregon.edu

ABSTRACT

As artificial intelligence becomes increasingly important across scientific disciplines, there is a growing need for scalable platforms that support training, knowledge sharing, and workforce development. The ICICLE AI Institute, funded by the National Science Foundation, focuses on democratizing artificial intelligence and cyberinfrastructure through open education, community engagement, and shared software infrastructure.

This paper presents the design and architecture of the ICICLE Training Catalog and demonstrates how the platform supports broader community engagement and workforce development through practical use cases and user personas. The catalog aggregates educational materials, tutorials, and technical documentation from multiple institutions into a centralized and structured repository.

We describe the website design, documentation design principles, and early usage evidence from privacy-preserving Umami analytics. We also discuss how the Training Catalog supports workforce development and how it can be leveraged by different user groups, including students, researchers, educators, and research software engineers. Finally, we highlight lessons learned from developing a collaborative training platform across institutions and discuss future directions for expanding its impact.

KEYWORDS

Artificial Intelligence, Cyberinfrastructure, Open Source, Training Material, Research Software

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page, along with the DOI information below. To copy otherwise, republish, post on servers or redistribute to lists requires prior specific permission from the first author.

© 2026 Journal of Computational Science Education
https://doi.org/10.22369/issn.2153-4136/17/2/7

1 INTRODUCTION

Artificial intelligence (AI) and cyberinfrastructure (CI) technologies are increasingly central to modern scientific discovery. The National Science Foundation-funded ICICLE AI Institute [1, 6] pursues new AI and CI technologies with the explicit goal of democratizing AI. The institute draws its challenges from three use-inspired domains—animal ecology, digital agriculture, and food access—and uses these challenges to develop software, tools, training materials, and workflows that make state-of-the-art AI technologies usable across scientific domains.

This work is accompanied by a growing need for accessible pathways through which researchers, educators, professionals, and students can learn to use AI tools and CI platforms effectively. Research computing communities commonly distribute training materials through learning management systems, workshop pages, single-project documentation sites, API portals, and institutional knowledge bases. Each of these channels is individually valuable, but each typically optimizes for a single audience or content type: a course site supports a scheduled or self-paced cohort, an API portal supports endpoint lookup, and a single-project documentation site supports users of a specific package. In multi-institutional AI and CI work, these separate channels can obscure how a software component, an educational module, an API, and a hosted service relate to one another.

To address this gap, ICICLE developed the ICICLE Training Catalog [2], which centralizes training resources produced by researchers, educators, and developers across the ICICLE ecosystem. By organizing these materials in a structured and searchable form, the platform broadens access to AI education and supports workforce development across institutions.

This paper describes the design and usage of the ICICLE Training Catalog. We present the platform's architecture, the conventions it uses to organize content, evidence of early adoption drawn from privacy-preserving analytics, and the ways in which it supports distinct user communities within the AI ecosystem.

2 WHAT IS ICICLE

ICICLE (Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment) [6] is a National Science Foundation (NSF) funded AI Institute focused on enabling the next generation of AI-powered cyberinfrastructure.

The institute brings together researchers, educators, and software engineers across universities, national laboratories, and computing centers. Its mission is to democratize artificial intelligence by making AI technologies accessible to scientists, educators, and students across disciplines.

ICICLE focuses on several key goals:

- Developing AI-powered cyberinfrastructure tools for scientific research
- Creating open educational resources for AI training
- Building reusable AI software components and workflows
- Supporting community-driven collaboration across institutions

To pursue these goals, the institute developed an open-source framework that makes materials findable for current users and approachable for new ones. We call the product the "ICICLE Training Catalog." The Training Catalog serves as a documentation source and as a pedagogical vehicle, connecting conceptual materials, hosted services, and reusable source code.

3 ICICLE TRAINING CATALOG

The ICICLE Training Catalog [2] is an open documentation platform that supports AI and CI education across a broad community of learners, educators, and researchers. It is designed as a bridge across the otherwise siloed channels that research computing communities use to distribute training material. The Training Catalog combines a documentation-site model for maintainability, the Diátaxis framework for consistent organization, metadata and thrust-based tags for cross-domain discovery, and privacy-preserving analytics for continuous improvement. Unlike a static list of workshop links or a single-project manual, it integrates software documentation, API specifications, educational materials, workshop resources, and source-code entry points into one searchable structure that supports workforce development in AI.

3.1 Building the Training Catalog

The Training Catalog is built using Docusaurus [5], a widely adopted open-source framework for creating documentation websites. Docusaurus enables content authoring in lightweight Markdown files, allowing contributors across institutions to easily create, maintain, and update documentation.

The architecture follows a static-site generation model with content version-controlled through repositories, enabling scalable and collaborative development. The platform integrates Umami [10], an open-source, privacy-preserving analytics solution deployed on TACC infrastructure via Tapis [9]. Umami captures aggregate engagement signals—page views, referrers, devices, browsers, operating systems, and navigation paths—without recording personally identifiable information. In Figure 1, the upper path depicts documentation ingestion and deployment; the Tapis Pods component (lower portion) hosts the Umami service and its PostgreSQL backend.

3.2 Structured Documentation with Diátaxis

To ensure consistency and usability, the ICICLE Training Catalog adopts the Diátaxis documentation framework [7], which organizes technical documentation into four user-centered categories:

- Tutorials
- How-to Guides
- Explanations
- References

Each documented resource includes a project overview together with one or more of these sections, enabling users to select content appropriate to their task (see Figure 2 for an example component page). The documentation also carries metadata such as last-updated timestamps, allowing users to assess the currency of the material.

Contributors can produce ICICLE-compatible README files through two complementary authoring pathways. The first is a repository-based template provided through the ICICLE GitHub workflow [4], which gives contributors a deterministic structure for project description, canonical tags, license information, references, acknowledgements, issue reporting, and Diátaxis sections. The second is an AI-assisted workflow in which contributors invoke the `icicle-readme-skill` [3] from AI-powered development environments such as Claude Code, Cursor, or Codex. In this mode, the skill guides the authoring or revision of a README so that it conforms to the ICICLE documentation standard while preserving project-specific technical content. Together, these pathways support both conventional template-driven authoring and interactive documentation development, lowering the barrier for contributors with different writing preferences and toolchains.

To operationalize this structure, a custom Python-based parser processes README files from individual projects and organizes them into Markdown files and directory structures compatible with Docusaurus. The parser extracts project metadata such as title, source repository, tags, last-updated information, and external service links; detects Diátaxis sections such as tutorials, how-to guides, explanations, and references; normalizes relative links and image paths; and emits Docusaurus-compatible pages and sidebar entries. When a README lacks required metadata or has ambiguous section headings, the parser reports warnings so maintainers can correct the source material before publication.

This automated pipeline ensures consistency across documentation while reducing manual effort. It also makes the Training Catalog easier to scale because contributors can continue to maintain project-level README files while the catalog build process produces a coherent public-facing site.

3.3 API Documentation

The platform's API documentation section presents OpenAPI v3 specifications for publicly available ICICLE endpoints, including Tapis-hosted services and project-specific APIs when component teams provide a specification. Integrated within Docusaurus, this section exposes endpoint descriptions, request and response schemas, authentication notes where applicable, and example calls. Examples are oriented toward common AI and cyberinfrastructure workflows, including command-line `curl`, Python, and JavaScript clients.

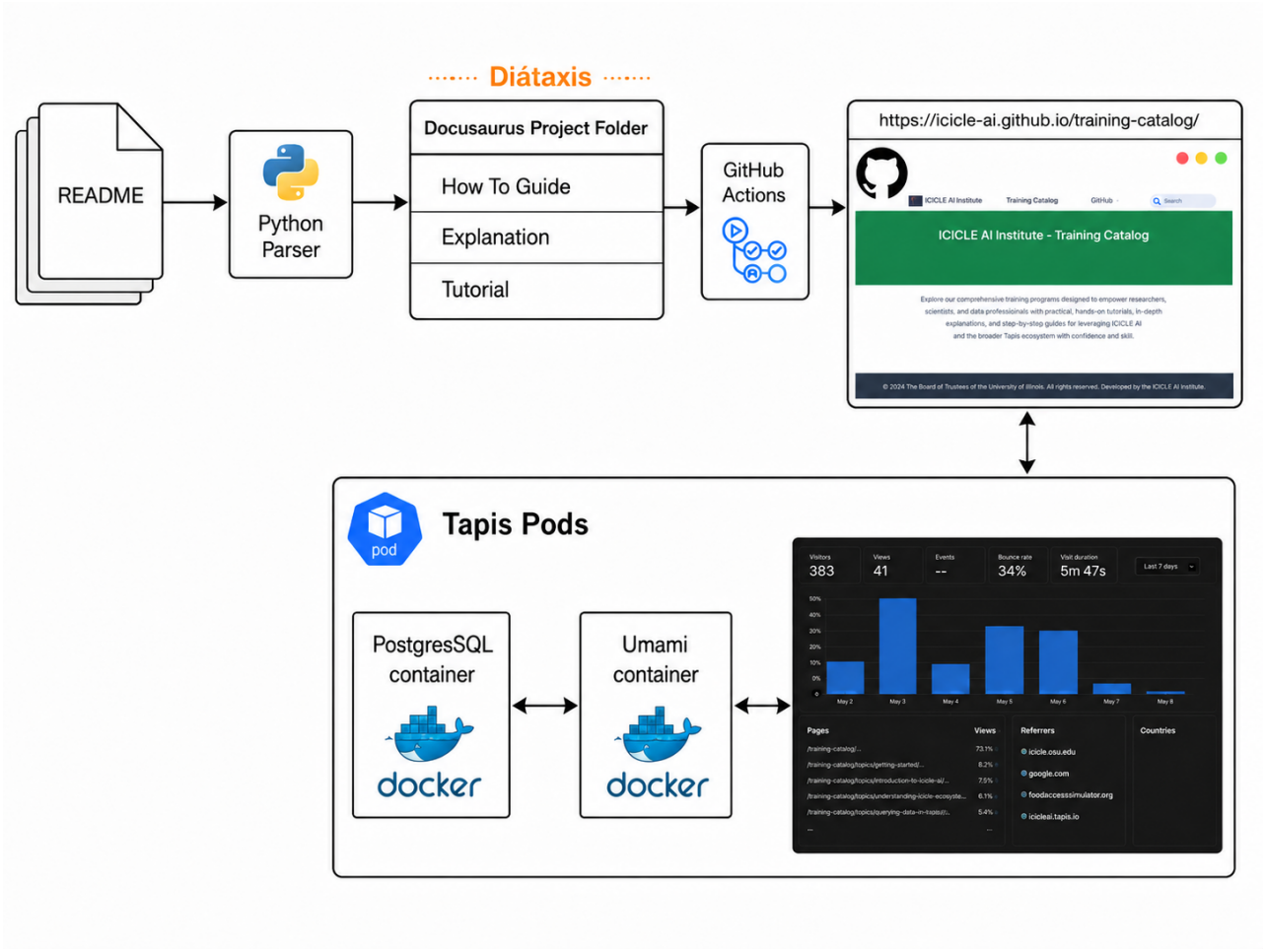


Figure 1: Architecture of the ICICLE Training Catalog.

3.4 Education Material and Workforce Development

The ICICLE AI Institute is committed to advancing AI and CI literacy through educational initiatives spanning K–12 education and professional development. The Training Catalog includes a dedicated education section that provides resources tailored to diverse learning levels and backgrounds.

These materials include concise "nutshell" videos that provide high-level overviews of activities across the ICICLE ecosystem, giving users a rapid orientation to key concepts and applications.

The platform additionally hosts educational documents that can be viewed inline through embedded interfaces or downloaded for offline and classroom use.

Educational tags such as *youth* and *video* support filtering and discovery, helping users locate materials matched to their level and interests.

3.5 Search and Thrust-Based Tagging

To enhance discoverability, the Training Catalog organizes content using thrust-based tags aligned with ICICLE's core research areas. Each resource is associated with one or more tags, allowing users to filter materials based on specific domains or topics.

The platform also integrates full-text, client-side search that enables keyword queries across all documentation pages. The search index is built during static-site generation and queried in the browser, eliminating the need for a hosted search service. The search interface highlights matches in titles and body content, enabling rapid identification of relevant pages.

3.6 Usage Analytics and Adoption

To characterize early adoption, we analyzed privacy-preserving Umami data collected from September 2025, when analytics collection was configured for the Training Catalog, through April 2026. The data comprise aggregate site metrics and CSV exports grouped by page path, referrer, device class, browser, and operating system.[1] The adoption analysis below emphasizes views, visits, visitors, and page-level navigation.

[1]In Umami, a *view* is a page load or client-side route change; a *visit* is a session containing one or more views; and a *visitor* is an approximate unique browser or device.

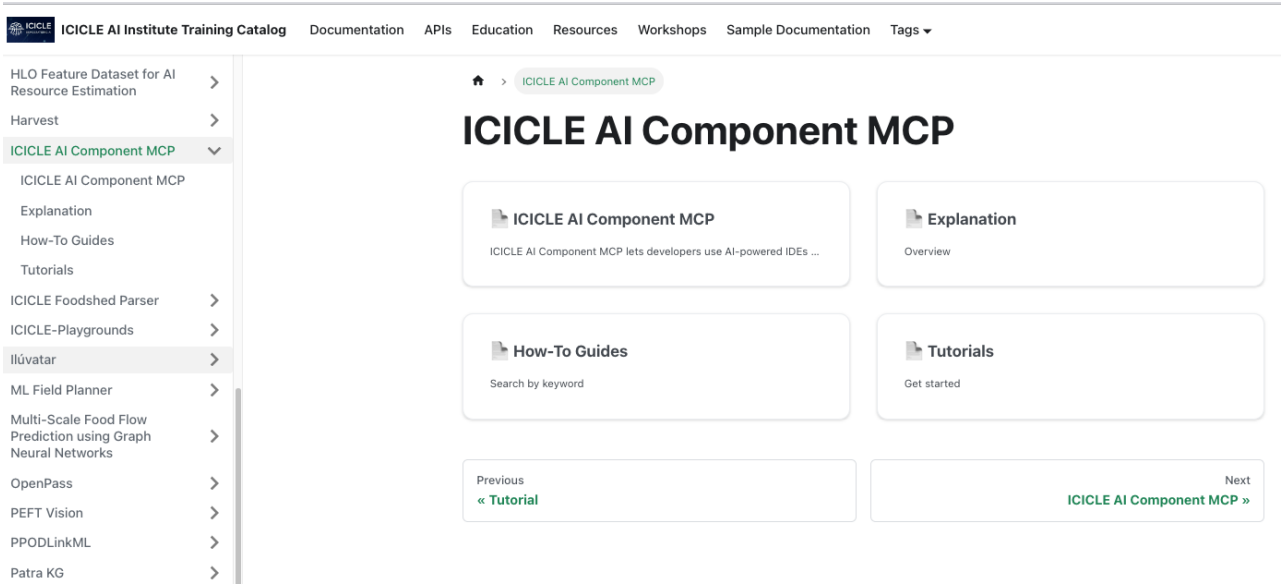


Figure 2: Component documentation page organized using the Diátaxis format.

Over the September 2025–April 2026 collection period, Umami recorded 2,690 views, 1,040 visits, and 244 visitors. Tables 1–5 summarize the resources accessed, the referral sources, and the client environments observed during this period. The per-attribute breakdowns do not sum to the overall view and visit totals: Umami records a row for an attribute (page path, referrer, device class, browser, or operating system) only when the value is identifiable from the request, and reports the top-ranked values for each category. The page-path export accounts for 1,862 of 2,690 views; the referrer export captures 169 visits with an identifiable referrer, of which the top entries shown in Table 2 account for 155; device class, browser, and operating system each cover 336 visits with a classified client environment.

The page-path export records, for each Training Catalog URL or route, the number of views it received; it does not count unique users. It contains 1,862 views distributed across 309 distinct paths. Grouping these paths by catalog area yields: documentation 1,273 views, education resources 158, the catalog home 139, API pages 137, workshops 63, and other or uncategorized paths 92. This distribution indicates that users routinely navigated beyond the landing page into the documentation, education, API, and workshop areas of the Training Catalog.

Table 1 reports the ten most-accessed catalog paths; the top five alone account for 18.0% of page-path views, with a long tail of project and education pages contributing smaller but non-trivial use. Referrers (Table 2) include ICICLE-owned entry points, general search engines, a domain-specific application site, Tapis-hosted infrastructure, and shared links from Microsoft Teams. Device classes (Table 3) indicate that the Training Catalog is used predominantly for large-screen technical work: laptop and desktop visits total 285 of 336 device-classified visits (84.8%), while mobile and tablet visits account for the remaining 51 (15.2%).

Table 1: Top ten catalog resources by view count.

Resource	Views
Catalog home	139
Documentation introduction	62
Other resources introduction	51
Education introduction	47
Workshops introduction	37
FEAST category	30
AI Ethics Curriculum	26
FEAST documentation	25
API introduction	23
ScienceAgent interface category	22

Table 2: Top referrers by visit count.

Referrer	Visits
icicle.osu.edu	52
google.com	28
foodaccesssimulator.org	19
icicleai.tapis.io	15
Microsoft Teams CDN	14
bing.com	11
linkedin.com	9
duckduckgo.com	7

Because the Training Catalog uses a self-hosted, privacy-preserving Umami deployment, these data do not include demographic attributes, institutional role, or individual identity. The deployment is also subject to common web-analytics limitations: browser tracking protections may undercount some visitors, automated traffic may not be completely filtered, and single-page-application routing can

Table 3: Device classes by visit count and share.

Device	Visits	Share
Laptop	202	60.1%
Desktop	83	24.7%
Mobile	46	13.7%
Tablet	5	1.5%

Table 4: Browsers by visit count.

Browser	Visits
Chrome	240
Safari	25
iOS browser	23
Firefox	19
Edge Chromium	16
Chrome iOS	5
iOS WebView	3
Opera	3

Table 5: Operating systems by visit count.

Operating system	Visits
Windows 10	164
Mac OS	92
iOS	32
Linux	21
Android OS	19
Windows 7	4
Chrome OS	3
Windows XP	1

affect how page transitions are recorded. We therefore treat the analytics as evidence of adoption and navigation behavior rather than as a causal evaluation of learning outcomes or demographic reach. Future evaluations will pair these passive analytics with voluntary surveys or workshop-specific feedback instruments that can measure how the learn–play–configure model affects user confidence, completion rates, and reuse of ICICLE software.

4 LEARN–PLAY–CONFIGURE PATHWAY

The ICICLE ecosystem serves users whose needs change as they move from initial exploration to reuse and extension. To support that progression, the Training Catalog organizes resources around a learn–play–configure pathway. Users *learn* by reading structured documentation, short-form videos, and tutorials; *play* by interacting with hosted services and examples; and *configure* by following repository links, cloning code, modifying configuration files, and adapting components for their own environments. Figure 3 illustrates this pathway with the FEAST Food Equity Access Simulation as a concrete example.

This pathway was used in practice by students in the San Diego Supercomputer Center Research Experience for High School Students Summer Program [8]. Students used the Training Catalog to identify ICICLE projects, move from project documentation to source repositories, download and install selected software, and run several projects on the SDSC Expanse system. Their use of the Training Catalog highlighted the importance of visible prerequisites, clear repository links, and compute-environment notes. These observations informed the Training Catalog's emphasis on metadata, how-to guides, and links between documentation pages and executable resources.

5 USER PERSONAS

The learn–play–configure pathway is intended to support several user groups without requiring separate sites for each audience.

5.1 Researchers

Researchers use the Training Catalog to discover AI tools, datasets, and workflows that can be applied to scientific problems. Thrust-based tags help them move from a domain such as food access or animal ecology to relevant components, while project pages connect explanations, tutorials, APIs, and source repositories. In the FEAST example, a researcher can move from the catalog entry to the hosted simulator and then to the source repository for adapting scenario parameters.

5.2 Students

Students benefit from structured learning resources that introduce AI and CI concepts through examples that can be run or inspected. Tutorials and workshop materials support first exposure, while links to hosted services and repositories let students move beyond static reading material. The SDSC REHS use case showed that explicit prerequisites and compute-environment notes are especially important when students transition from reading documentation to running code.

5.3 Educators

Educators use the Training Catalog to find teaching materials and incorporate AI-related modules into their courses. The structured documentation format helps instructors adapt resources for classroom environments and connect students to runnable examples beyond static reading material.

5.4 Research Software Engineers

Research software engineers (RSEs) are both users and contributors. As users, they rely on the Training Catalog to understand how components are documented, built, and deployed. As contributors, they use the README template, AI-assisted `icicle-readme-skill` workflow, parser feedback, and metadata conventions to prepare project documentation that can be ingested into the catalog. This role connects the user-facing documentation experience to the maintenance work required to keep component pages current.

6 LESSONS LEARNED

Several lessons emerged from developing the Training Catalog across institutions. First, a shared documentation framework is a coordination tool, not only a presentation choice. Diátaxis gave contributors a common vocabulary for deciding whether material should teach a first-time workflow, answer a specific task question, explain concepts, or provide reference details.

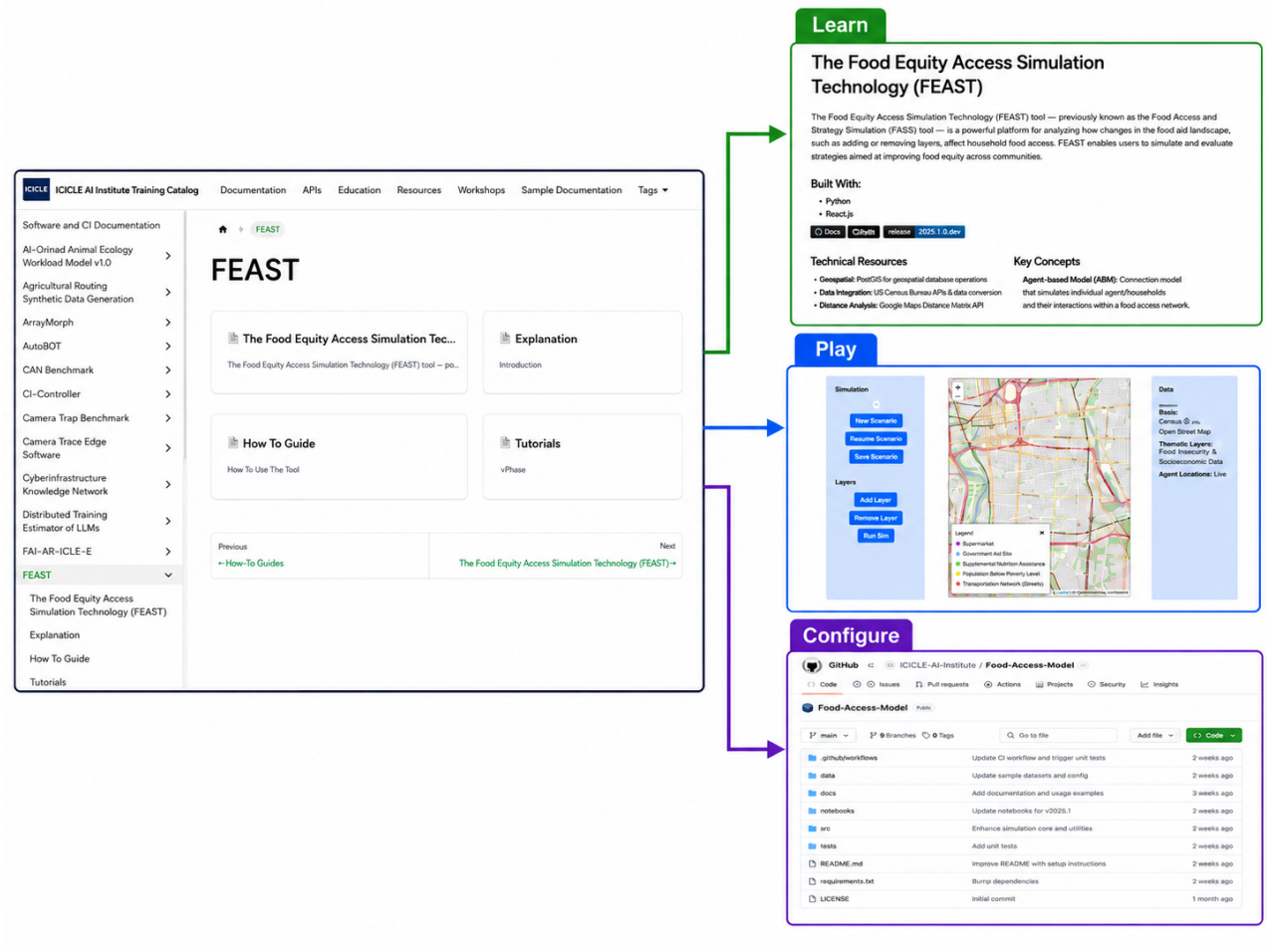


Figure 3: Learn–play–configure pathway, illustrated with FEAST.

Second, automation needs an explicit metadata contract. The Python parser reduced manual formatting work, but only after the team agreed on required fields, section names, link conventions, and update timestamps. Without that contract, automated ingestion can reproduce inconsistencies at larger scale. In early onboarding attempts, parser warnings often pointed to practical problems such as missing canonical tags, ambiguous section headings, and repository links that had drifted from the current project layout.

Third, repository-based documentation reduced cross-institutional friction. Docusaurus, GitHub Actions, and Markdown allowed contributors from different institutions to work in familiar tools while still producing one public catalog. A recurring friction point was that contributors had different local documentation habits: some maintained detailed README files, some stored installation notes in workshop slides, and some assumed users already knew the relevant compute environment. The repository-based workflow made these differences visible and easier to resolve through review.

Fourth, privacy-preserving analytics are useful but incomplete. Umami provided enough information to identify popular resources, referrers, and device patterns, but it did not answer demographic questions or measure learning outcomes. We therefore plan to pair analytics with voluntary feedback mechanisms when evaluating training impact.

Fifth, links between documentation and executable resources require active maintenance. Hosted services, repositories, and workshop materials can change independently across institutions. Dead links and outdated setup instructions were among the most visible failure modes for users, so the Training Catalog needs regular link checks and ownership metadata in addition to initial onboarding documentation.

7 CONCLUSION

The ICICLE Training Catalog serves as a central hub for the broader AI and cyberinfrastructure community, providing a single, structured entry point into the full range of resources produced across the ICICLE ecosystem. By aggregating research software documentation, API specifications, educational materials, workshop content, and multimedia resources into one discoverable platform, the Training Catalog lowers the barrier for external audiences—students,

researchers, educators, and practitioners—to explore and engage with ICICLE's work.

The platform's design combines the Diátaxis documentation framework with thrust-based tagging, full-text search, and the learn–play–configure interaction model to support users with varying levels of expertise and intent. Whether a user is encountering ICICLE tools for the first time or integrating them into active research workflows, the Training Catalog provides the structure and navigation needed to find relevant content quickly.

Early Umami analytics show measurable adoption: from September 2025 through April 2026, the Training Catalog recorded 2,690 views, 1,040 visits, and 244 visitors, while the page-path CSV export showed engagement across documentation, education, API, and workshop pages. These metrics are an initial baseline rather than a complete evaluation, but they provide a foundation for improving content organization and identifying resources that need clearer onboarding paths.

Looking ahead, the Training Catalog will continue to evolve alongside the ICICLE community. Planned improvements include richer filtering by compute requirements and experience level, publication of the documentation parser and validation workflow, more explicit compute-environment guidance for hands-on examples, and voluntary feedback instruments that can complement privacy-preserving analytics with learning-outcome data.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Award No. OAC-2112606 through the ICICLE AI Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors would like to thank the ICICLE community and partner institutions for their contributions to the development of open AI cyberinfrastructure and training resources.

The authors used Claude (Anthropic) to assist with spell-checking, grammar, and minor writing corrections. All technical content, analysis, and conclusions are solely the work of the authors.

REFERENCES

[1] ICICLE AI Institute. 2021. ICICLE: Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment. https://icicle.ai. Accessed: 2026-03-25.

[2] ICICLE AI Institute. 2024. ICICLE Training Catalog. https://icicle-ai.github.io/training-catalog/. Accessed: 2026-03-25.

[3] ICICLE AI Institute. 2025. ICICLE README Skill. https://github.com/ICICLE-ai/training-catalog/blob/docusaurus-demo/icicle-readme-skill/SKILL.md. Accessed: 2026-05-18.

[4] ICICLE AI Institute. 2025. ICICLE README Template. https://github.com/ICICLE-ai/training-catalog/blob/docusaurus-demo/icicle-readme-skill/templates/README.md. Accessed: 2026-05-18.

[5] Meta Open Source. 2024. Docusaurus. https://docusaurus.io/docs. Accessed: 2026-03-25.

[6] Dhabaleswar K. Panda, Vipin Chaudhary, Eric Fosler-Lussier, Raghu Machiraju, Amit Majumdar, Beth Plale, Rajiv Ramnath, Ponnuswamy Sadayappan, Neelima Savardekar, and Karen Tomko. 2024. Creating intelligent cyberinfrastructure for democratizing AI. *AI Magazine* 45, 1 (2024), 22–28. https://doi.org/10.1002/aaai.12166

[7] Daniele Procida. 2021. Diátaxis: A systematic approach to technical documentation authorship. https://diataxis.fr/. Accessed: 2026-03-25.

[8] San Diego Supercomputer Center. n.d.. UCSD Research Experience for High School Students (REHS). https://education.sdsc.edu/studenttech/rehs/. Accessed: 2026-03-27.

[9] Joe Stubbs, Richard Cardone, Mike Packard, Anagha Jamthe, Smruti Padhy, Steve Terry, Julia Looney, Joseph Meiring, Steve Black, Maytal Dahan, Sean Cleveland, and Gwen Jacobs. 2021. Tapis: An API Platform for Reproducible, Distributed Computational Research. In *Advances in Information and Communication*, Kohei Arai (Ed.). Springer International Publishing, Cham, 878–900.

[10] Umami Software, Inc. 2022. Umami Analytics. https://umami.is/. Accessed: 2026-03-25.