# Metrics for Gateways: Why and How

Jeanette Sperhac
*San Diego Supercomputer Center*
*University of California, San Diego*
La Jolla, CA, USA
ORCID# 0000-0003-0436-9318
jsperhac@sdsc.edu

Richard K. Wellner, Jr.
*San Diego Supercomputer Center*
*University of California, San Diego*
La Jolla, CA, USA
rkw@sdsc.edu

Amit Chourasia
*San Diego Supercomputer Center*
*University of California, San Diego*
La Jolla, CA, USA
amit@sdsc.edu

***Abstract*—Science Gateways, which provide researchers, educators, and students with streamlined access to advanced computational resources, have become increasingly important to the scientific community. Though not central to the typical gateway's mission, streamlined collection and reporting of usage statistics and other metrics is an important part of sustaining a healthy gateway. Metrics describing usage and user activity enable gateway managers and owners to plan for expansion, improve services to the user base, and set and evaluate actionable goals. We will discuss these and other motivations for collecting gateway metrics, along with methods of data collection, processing, and reporting. We will describe the HUBzero metrics collection method, which has been robust and extensible enough to serve its gateway platform for the last 20 years, and propose enhancements for gateway metrics going forward.**

***Index Terms*—metrics, usage, statistics, gateways**

## I. Introduction

Science gateways have become common engines for furthering scientific practice. These portals offer web-based user interfaces that provide academic or research communities with curated applications, whether computational, data-intensive, or collaborative. [1], [2] Gateways are tailored to the requirements of their communities, supporting such diverse features as access to high performance computing, large data repositories, computational tools, and educational materials such as whitepapers, publications, slidesets, and videos.

Science Gateways must frequently defend their existence to their host institutions and funding agencies. This is a natural process, as institutions want to support resources that provide positive impact to research, publication, and teaching. Gateway usage data detailing user registrations, computations run, citations, document downloads, and similar statistics help build a case for the importance of a gateway in its community.

For these reasons, consideration of metrics collection and reporting should be an early concern when creating a new gateway. Registration of new users, releases of new computational tools, gateway use for coursework, and other figures can all be captured by metrics. These data then help sustain the gateway by demonstrating its usefulness to its community.

In this paper we motivate and discuss useful gateway metrics, and considerations for collecting, aggregating, and presenting them. We will look at the features of the standard HUBzero metrics package, which has provided metrics for HUBzero gateways for 20 years. [3] We will also discuss caveats and recent enhancements and extensions to this package, and consider improvements for the future.

This work is supported by National Science Foundation award EEC #1227110, (Network for Computational Nanotechnology Cyber Platform).

## II. Why collect metrics?

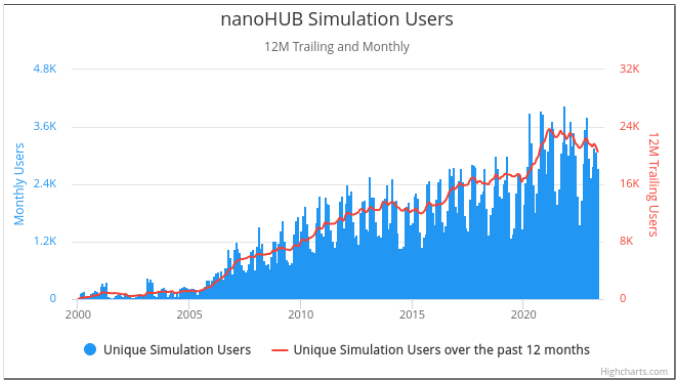


Fig. 1. Simulation user counts for the NanoHUB gateway are shown, for monthly and 12-month trailing periods.

Collecting gateway metrics can be useful for numerous audiences and for a host of reasons. Information about usage can help gateway owners and PIs decide on features to expand or retire by bringing focus to both well-used and underutilized areas of the gateway. This knowledge can help establish whether gateway infrastructure, including software and physical equipment, is sufficient to meet community demand. It can also help gateway owners identify new audiences to which they might promote their gateway. Furthermore, metrics with unusual values can help identify cases where gateway resources are failing, misconfigured, or being misused or hacked by bad actors.

In addition to gateway owners, there are others who benefit from the availability of metrics. First, gateway users: Metrics can help attract users by demonstrating that a gateway's resources are both current and commonly accessed. Gateway resource contributors profit when metrics show how widely their tools or publications have been accessed, thus establishing the reach of their work. Finally, and quite importantly, metrics help assure funders that resources allocated to a gateway are being sensibly used to the benefit of the gateway community.

Gateway owners' goals and targets can include counts of new user registrations, workshop enrollments, data sources, or computational tool releases. After these objectives have been identified, metrics collected during regular operation can help gateway owners evaluate whether they are meeting their goals. [4], [5] Metrics are important; they can identify whether the gateway is achieving success.

## III. Case Study: HUBzero metrics package

### A. The HUBzero platform

The HUBzero Platform for Scientific Collaboration is a proven open-source platform that has served the science gateways community for 20 years. It enables gateway users to run software tools and computations directly in their web browser, without having to compile or install code on local systems. [3] The platform also lets users publish and share resources such as datasets, online presentations, whitepapers, and other materials. Additional features encourage collaboration among Hub users with communication tools such as online groups, forums, and blogs. HUBzero's capabilities are evident in the platform's flagship gateway, NanoHUB (nanohub.org), which saw 2,074,000 unique visitors during 2021. NanoHUB users can run computations with its more than 800 software tools and 7000 non-tool resources, including online presentations and papers. Hands-on class sessions and workshops around the world use NanoHUB's features for instruction. [6]–[8]

The HUBzero gateway's architecture consists of a database server and webserver, with optional execution hosts where software containers run computational tools. Middleware coordinates the container sessions with user sessions. The webserver handles the gateway's user interactions, including registering and authenticating users, controlling access to hosted resources such as tools and data, and managing tool and resource development workflows. [9], [10]

Hosted computational tools, also known as simulations, are found at the heart of many gateways. When a user launches a tool on a gateway, the gateway's execution host starts a virtual container. These tool containers are configured to handle specific computational needs, providing the needed software libraries, memory, and disk space. Tools that require specialized computational resources or parallel execution can be configured to submit jobs to high performance computing resources. All tool run data, including user interaction time, CPU and disk use, and remote computation, are minutely logged by the gateway for later analysis. User access to non-tool resources is similarly logged so that access to papers, presentations, and datasets can also be evaluated. [11]

### B. HUBzero and metrics

The long-lived HUBzero platform has made metrics collection and presentation a standard practice for its gateways, known as Hubs. There are many opportunities for collecting usage information on the typical gateway. From logging casual page visits, to capturing user registration information, to aggregating data on tool or resource usage, data are collected on most user interactions with the HUBzero platform.

Nightly HUBzero metrics processing utilizes logs from the gateway's webserver and content management system as sources of input. These log data are parsed, loaded into relational databases, and combined with tool run data. Spurious and suspected spam records are discarded. Web hits from the logs are matched to existing resources deployed on the Hub to assemble counts of resource accesses by users. User location information is inferred from IP addresses for reporting in the aggregate; other user information, such as academic affiliations, are extracted from user profiles and aggregated. Users have the option to provide user profile information that describes their background and demographics, summarizes their academic or professional achievements, and more.

Metrics processing prepares a wide array of usage data on areas of the gateway for which user-facing reports have not yet been developed. This means that untapped metrics data sources are available for the development of new reports. On the other hand, some usage data are not suitable for end user reporting because of sensitivity (e.g. inferred user location or demographic information), and are only reported in aggregate. No sensitive information, nor individual user data, is reported in HUBzero metrics.

## IV. HUBzero Metrics reporting

When rich data sources on gateway usage are available, there are many possibilities for presenting usage metrics. Usage reports can be tailored to the casual visitor, the individual contributing user, the gateway administrator, and other audiences. On the HUBzero platform, contributing users can refer to their own dashboard pages to view usage data for materials they have authored, such as computational tools and publications. This helps gateway members easily determine how many unique users and runs their contributions have garnered (see Section IV-A). Specific usage data on individual computational tools is also tabulated for end-user use (see Section IV-B). The Hub's general usage summary shows visitor and registered user counts and counts of available tools and resources, along with aggregated data about users and their engagement with computational tools (see Section IV-C).

### A. Contributor View of Usage

The HUBzero gateway publication model relies on the initiative of gateway users to contribute computational tools and other resources to the community. Accordingly, the Hub provides contributing users with a snapshot of usage reporting on their contributions. Thus, gateway members can view the usage statistics describing their contributions on their own user profile page. These tabulated numbers are updated with each run of metrics processing.

Three tables shown on the user profile page metrics tab provide this information. For some Hub called *hubname*, and some *member_id*, the tab is found at https://*hubname*/members/*member_id*/usage. In the tab, a usage overview (Table I) summarizes the user's contributions, including citation count and course usage:

| HUBzero Contributor Metrics Overview |
|---|
| Count of contributions |
| Rank by Count of contributions |
| Date of First Contribution |
| Citation count on Contributions |
| Usage in Courses/Classrooms |

TABLE I
HUBZERO CONTRIBUTOR METRICS OVERVIEW

The contributor's simulation tool usage display (Table II) shows current usage for each of the user's contributed tools, including total runs and citations and users served. Figure 2 shows a section of this report for a specific NanoHUB contributor. The report provides links to tool pages and to overall tool statistics (see Section IV-C3):

| Metrics reported for each contributed tool |
|---|
| Users served in last 12 months |
| Simulation Runs in last 12 months |
| Total users served |
| Total Simulation Runs |
| Citation Count |
| Published Date |

TABLE II
HUBZERO CONTRIBUTOR SIMULATION TOOL USAGE

Table 2: Simulation Tool Usage

| # | Tool Name | Users served in last 12 months | Simulation Runs in last 12 months | Total users served | Total Simulation Runs | Citations | Published On |
|---|---|---|---|---|---|---|---|
| 1 | Jupyter Lab (201904) | 8 | 22 | 8 | 22 | - | 08 Feb 2023 |
| 2 | Jupyter Lab (202105) | 16 | 112 | 16 | 112 | - | 08 Feb 2023 |
| 3 | Silvaco TCAD | 640 | 5,347 | 640 | 5,347 | - | 27 Sep 2022 |

Fig. 2. Contributor's tool metrics demonstrate the 12-month and lifetime impact of a Hub user's contributions.

The contributor's non-tool usage display ("and more" usage, Table III) shows metrics for each of the non-tool resources they have authored, such as notes, papers, videos, or slides:

| Metrics reported for each non-tool resource |
|---|
| Users served in last 12 months |
| Total users served |
| Citation Count |
| Published Date |

TABLE III
HUBZERO CONTRIBUTOR NON-TOOL RESOURCE USAGE

Contributor report functionality is implemented via a HUBzero core system plugin. This code is available on all HUBzero instances, and relies on nightly metrics processing. to populate the tables it queries.

*1) Features and Caveats:* These reports provide a convenient mechanism for contributing users to determine the impact their work has on the gateway community. Users can tout the current and lifetime usage statistics of their contributions for tenure committees and grant proposals.

Despite their usefulness, several caveats for contributor reports can be noted. First, the report displays the overall and current month's total usage for each contribution, but no historical monthly data is included (See Figure 2). Second, the contributor reports are tied only to the user who deployed the resource on the Hub. Reports do not include any associated co-authors, who were also responsible for its development and creation. Furthermore, the report does not retrieve metrics for specific time periods, create plots or data visualizations, or provide data download. Adding plotting, historical data, and download features, as well as making the report available to collaborators and coauthors, would enhance this report.

### B. Individual Tool View of Usage

For many Hubs, computational tools are central; some gateways support hundreds of such tools. Each deployed tool on a HUBzero gateway features its own usage report, available from its landing page. For a Hub called *hubname*, the *toolname* usage report is found at https://*hubname*/resources/*toolname*/usage. Timeseries plots with adjustable date ranges and zoom controls, and a variety of tabulated data, are displayed in this report (Table IV):

| Tool usage report |
|---|
| World usage (visualized on map, NanoHUB only) |
| Cumulative Simulation User count, as timeseries plot |
| Users by organization type |
| Users by country of residence |
| Cumulative Simulation Runs, as timeseries plot |
| Wall Clock Time |
| CPU Time |
| User Interaction Time |

TABLE IV
HUBZERO INDIVIDUAL TOOL USAGE

Two plots are featured in the report. Figure 3 shows the timeseries plot of tool users for the *Multispec* tool deployed on MyGeoHub (mygeohub.org). [12] The tool's cumulative user counts are shown for each month from 2014-10 to 2023-04. The x-axis shows the month and year; the y-axis shows the user count. Also displayed is part of the chart aggregating the tool's users by their organization type (e.g. academic, industry, or other) and location.

Figure 4 shows this report's timeseries plot of tool runs for the *PN Junction Lab* tool deployed on NanoHUB. The tool's cumulative simulation run counts are shown for each month from 2020-08 to 2023-05. The x-axis shows the month and year; the y-axis shows the run count. The total run count, and average and total wall clock time, CPU time, and user interaction time logged for this tool are also tabulated.

Tool metrics report functionality is implemented via a HUBzero core system plugin. This code is available on all HUBzero instances, and relies on nightly metrics processing

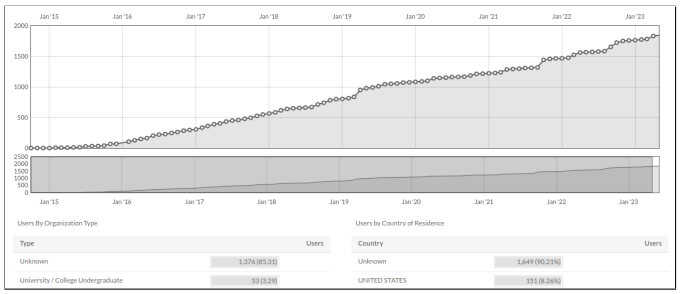


Fig. 3. Timeseries plot of cumulative simulation user count, MyGeoHub's *Multispec* tool, from 2014-10 to 2023-04.

to populate the tables it queries. The usage metrics shown in the page are updated at the close of each month.

*1) Features and Caveats:* The tool usage reports serve several purposes for the Hub community. They enable tool and Hub administrators to assess the impact of each tool deployed on the gateway, helping them to appropriately prioritize maintenance, upgrades, and release schedules. The tool's usage history can be referenced on grant proposals and reviewed by prospective users, establishing the tool's credibility and reach. Some sense of the audience for the tool can be determined by considering the percentage of users associated with different organization types and physical locations.

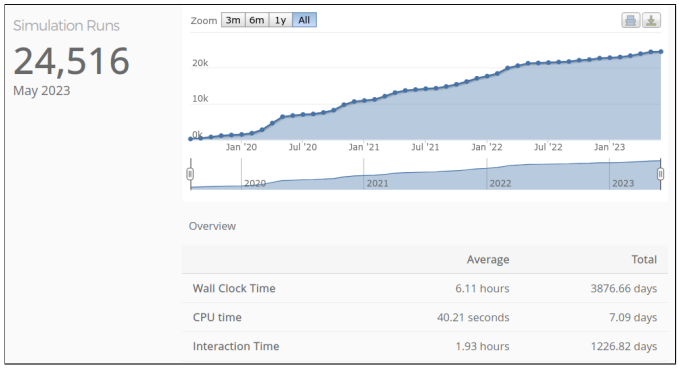


Fig. 4. Timeseries plot of cumulative simulation jobs, NanoHUB's *PN Junction Lab* tool, from 2020-08 to 2023-05.

There are also several caveats associated with tool metrics reports. First, while the report presents aggregated user organization type and location, these metrics typically display a high percentage of unknown values. Though organization type is limited by its reliance on user self-reporting, the user's location is determined by IP. The code used for this calculation had some limitations and has recently been improved. Rollout of this change should improve the data presented. On NanoHUB, the report also includes a visualization of the worldwide usage of a tool, indicated by the location of pins on a world map (not shown). This visualization could be improved by using color gradations instead of individual pins to indicate user counts per country or region.

Finally, the plots should be improved by clarifying the plot, axis, and data labels, providing mouseovers with additional information, and making zoom controls more intuitive. Plot titles and mouseovers should indicate whether the counts displayed are cumulative, yearly, or monthly. Adding download of the raw data in a convenient format would enable offline visualization or analysis. Finally, while plot download functionality is available on NanoHUB, it is not available on all Hubs. Adding and standardizing these features would enhance the tool usage report.

### C. Summary View of Usage

A prominent page on each HUBzero gateway summarizes its overall usage statistics. For a Hub called *hubname*, it can be found at https://*hubname*/usage.

*1) NanoHUB Usage Summary:* In the specific case of the NanoHUB gateway, a custom usage page enables users to interact with timeseries plots of different usage metrics. Each plot displays the monthly count and cumulative 12 month trailing count for one of the following metrics:

- visitors (non-registered user)
- registered users
- published resources
- published computational tools
- simulation users (Figure 1)

For example, Figure 1 shows NanoHUB's unique simulation user count: the monthly count is shown by blue bars, and the cumulative count computed for the last 12 months is shown by a red line.

*2) General Hub Usage Summary:* For most HUBzero gateways, the summary page shows user, visitor, and tool usage plots. In the metrics summary page, user visit and resource download counts are overlaid in a timeseries plot summarizing user activity. These are accompanied by a table aggregating users by organization (self-reported) and country of residence (inferred by IP). Hub administrators have the option to add this usage summary page to a gateway menu.

Tool usage reporting is summed for all the gateway's tool offerings. Timeseries plots of total tool user counts and simulation job (tool run) counts per month are shown. Gateway-wide tool usage resource consumption figures are displayed in tabular form, including Total CPU Time; Total Wall Time; and Total Interaction Time. Also shown in tabular form are cumulative totals for user tool interactions:

- Average Number of Simulation Jobs per User
- Average Time between User's First and Last Simulation
- Count of Users with more than 10 min of CPU time
- Count of Repeat Users running more than 10 Jobs
- Count of Repeat Users over more than 3 months duration

*3) Tools usage tab:* The tools usage tab is available from the Hub usage summary page for all Hubs. It shows the gateway's top 10 tools by several different criteria:

- User Ranking
- Simulation User Count
- Simulation Job Count
- Wall Time
- CPU Time

- User Interaction Time; and
- Citations

This functionality is implemented via a HUBzero core system plugin associated with tools usage. It is available on all HUBzero instances and relies on nightly metrics processing to populate the tables it uses for reporting. The usage metrics shown in the page are updated at the close of each month.

### *D. Metrics caveats and lessons learned*

In the HUBzero metrics implementation, various features can be improved by further development, refinements, and fixes. Clarity and usability of reporting is one area, and is most visible to the end user. Improvements in plot labeling, and supplying clear definitions and explanations for the metrics shown, would help clarify reporting (see Sections IV-A and IV-B above). Enabling raw data download and chart export would help users who wanted to perform additional offline analyses of usage data. The enhancements made to NanoHUB's summary Usage page provide a cleaner and more informative report than the default report provided for general-case Hubs (see Section IV-C and Figure 1). These enhancements should be added to the general Hub usage page, and further improved by data and chart export. Additionally, while the tools usage tab offers interesting criteria, its tabular format seems like a lost opportunity. (See IV-C3). Offering interactive plotting and direct comparison of these metrics, and download of the data, would be informative.

Scaling presents another set of problems for metrics. A successful gateway may grow until its user and job count outstrips its existing metrics processing, whether in terms of storage or processing time. Once nightly processing spans more than 12 hours, for example, metrics processing can begin to affect normal usage of the gateway. As the largest HUBzero gateway, NanoHUB, has grown, it has required periodic refactoring of metrics processing to rein in the execution time. Implementing queueing, data table indexing, and optimizing query performance has helped to scale nightly processing to fit the available overnight window.

Technical debt can also accumulate in a long-lived system. As time passes and changes are made, bugs accumulate and libraries need updates, and parts of the system may no longer serve as they once did. Occasional review is necessary to ensure metrics still work as designed, which is why we have recently undertaken work to review and improve the standard Hub metrics processing. These fixes and updates will provide better, more timely, and more robust metrics to the Hubs.

Spam content and bots present a problem for metrics collection. Since Hub access by bots skews true usage counts, logs must be inspected for bot activity prior to metrics processing. Meanwhile, collecting metrics on spam content wastes processing time. Maintaining a current list of known bots is important for scouring spurious Hub accesses from log data, and regularly scrubbing junk content helps keep the Hub from wasting processing on faux resources. Logging and monitoring of nightly data processing for metrics can help to indicate these and other processing problems in a timely manner.

## V. Future Work

The collection and presentation of metrics has evolved alongside the gateways they describe. Bringing new resources and communities online challenges us to better represent the features and usage of our systems for different audiences. As the platform's metrics evolve, they must consider gateway community needs, the scalability of processing, usability of reports, and metrics processing runtimes.

As the HUBzero platform and its communities continue to evolve, usage reporting should be expanded to include additional parts of the system that receive less attention in a tool-centric gateway ecosystem. Reporting could encompass usage information for remote data sources, storage, or computational sources, and for individual non-tool resources, which presently receive short shrift. Collaborative features, such as forums, blogs, and groups, could be better featured in usage reporting, to help guide the priorities of Hub owners. Administrator-level reporting on especially active or prolific users could assist Hub owners in running their online community.

The general HUBzero metrics collection package has recently been reviewed for code fitness, scalability, algorithms, library upgrades, and other considerations. Once these improvements are in place, all Hubs will benefit from improved metrics collection and processing.

New approaches can also be leveraged to improve gateways metrics collection and presentation. For example, open-source packages such as Grafana [13] and Prometheus [14] are used by data centers and operations teams for data collection, visualization, and reporting. Such packages are under consideration for future metrics offerings.

## VI. Conclusion

Gateway owners, users, and others benefit from metrics collection. Metrics can be used to track growth in the number of accounts, the usage of deployed resources, the length or resource consumption of sessions, the populations using the resources, and so forth. They can be used to establish the strength of a gateway community, to chart progress on goals, to plan for future expansion, to bolster proposals to funding agencies, and to establish the need for resources. In this paper we have explored the utility of gateway metrics and an example of metrics collection, processing, and presentation.

## Acknowledgment

The authors would like to thank Michael Zentner and the entire HUBzero team, including Dave Benham, Juliana Casavan, Steve Clark, Gene Eberhardt, Sandra Gesing, Nick Kisseberth, Daniel Mejia, Pascal Meunier, Subhash Ramesh, Ilya Shunko, Jack Smith, Claire Stirm, Danielle Whitehair, Mona Wong, Jesse Woo, Choonhan Youn, and Mark Zhuang.

We also thank the NanoHUB and MyGeoHub teams, and all of our Hub owners, for providing actionable feedback and evolving requirements for metrics.

## REFERENCES


[1] "Science Gateways Institute and SGX3 (SGCI/SGX3)," https://sciencegateways.org. Accessed June 9, 2023.

[2] Wikipedia, https://en.wikipedia.org/wiki/Science_gateway. Accessed June 9, 2023.

[3] M. Mclennan and R. Kennell, "HUBzero: A platform for dissemination and collaboration in computational science and engineering," *Computing in Science and Engineering*, vol. 12, no. 2, pp. 48–52, 2010.

[4] N. Wilkins-Diehr, "Measuring success: How science gateways define impact." San Diego, CA: Science Gateways 2019, October 2019. [Online]. Available: osf.io/tkzuy

[5] P. Calyam, N. Wilkins-Diehr, M. Miller, E. H. Brookes, R. Arora, A. Chourasia, D. M. Jennewein, V. Nandigam, M. Drew LaMar, S. B. Cleveland, G. Newman, S. Wang, I. Zaslavsky, M. A. Cianfrocco, K. Ellett, D. Tarboton, K. G. Jeffery, Z. Zhao, J. González-Aranda, M. J. Perri, G. Tucker, L. Candela, T. Kiss, and S. Gesing, "Measuring success for a future vision: Defining impact in science gateways/virtual research environments," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 19, p. e6099, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6099

[6] G. Klimeck, M. McLennan, S. P. Brophy, G. B. Adams III, and M. S. Lundstrom, "NanoHUB.org: Advancing Education and Research in Nanotechnology," *Computing in Science and Engg.*, vol. 10, no. 5, p. 17–23, September 2008. [Online]. Available: https://doi.org/10.1109/MCSE.2008.120

[7] G. Klimeck, L. K. Zentner, K. P. C. Madhavan, V. A. Farnsworth, and M. Lundstrom, "Network for computational nanotechnology - a strategic plan for global knowledge transfer in research and education," in *2011 IEEE Nanotechnology Materials and Devices Conference*, 2011, pp. 75–79.

[8] K. Madhavan, L. Zentner, V. Farnsworth, S. Shivarajapura, M. Zentner, N. Denny, and G. Klimeck, "nanoHUB.org: cloud-based services for nanoscale modeling, simulation, and education," *Nanotechnology Reviews*, vol. 2, no. 1, pp. 107–117, 2013. [Online]. Available: https://doi.org/10.1515/ntrev-2012-0043

[9] N. Denny, M. Zentner, and G. Klimeck, "Visualizing User Interactions with Simulation Tools." Austin, TX: Science Gateways 2018, September 2018. [Online]. Available: https://doi.org/10.6084/m9.figshare.7228274.v3

[10] E. Huebner, "From Linux Desktop Tools to Jupyter Notebooks to web-based widgets: Deploying your application on the Hubzero Platform." Austin, TX: Science Gateways 2018, September 2018. [Online]. Available: https://doi.org/10.6084/m9.figshare.7228274.v3

[11] D. McKay, M. Zentner, and G. Klimeck, "Clustering Download Events to Identify Classrooms." Austin, TX: Science Gateways 2018, September 2018. [Online]. Available: https://doi.org/10.6084/m9.figshare.7068179.v2

[12] R. Kalyanam, L. Zhao, C. Song, L. Biehl, D. Kearney, I. L. Kim, J. Shin, N. Villoria, and V. Merwade, "Mygeohub—a sustainable and evolving geospatial science gateway," *Future Generation Computer Systems*, vol. 94, pp. 820–832, 2019. [Online]. Available: https://doi.org/10.1016/j.future.2018.02.005

[13] Grafana Labs: https://grafana.com/.

[14] B. Rabenstein and J. Volz, "Prometheus: A next-generation monitoring system (talk)." Dublin: USENIX Association, May 2015.