CFDE Metrics Handbook

2024-08-14

Table of contents

Preface

The Common Fund Data Ecosystem (CFDE) is a critical infrastructure supporting data generated by NIH Common Fund programs. As the volume of biomedical data grows, there is an increasing need for systems that not only store data but also facilitate its effective use by researchers, clinicians, and policymakers.

This report examines the metrics that can quantify the impact, significance, value, and costs of the CFDE. The metrics listed in the Metrics section valuating the ecosystem's contributions, including its role in advancing scientific research and collaboration, as well as its broader economic and social effects.

The proposed framework includes both qualitative and quantitative assessments to address the challenges of measuring a complex and evolving system like the CFDE. By establishing clear metrics, this report aims to provide a basis for evaluating the performance of the CFDE and informing decisions about its future development and sustainability.

Why?

The collection of metrics is expected to address:

1. Aid in the development of a sustainability planning for CFDE components.
2. Need for NIH to have a clear understanding of the impact, signficance, and costs of the CFDE.
3. Need for DCCs to demonstrate the value of their contributions to the CFDE through reporting to the NIH.
4. Need for the CFDE to demonstrate its value to the broader scientific community.
5. Need for the CFDE to identify areas for improvement and optimization.
6. Aid in decision-making and resource allocation within the CFDE.
7. Provide a basis for continuous improvement and evolution of the CFDE.

1 Operational Concerns

1.1 Federated Data Collection

Metrics and evaluation within the Common Fund Data Ecosystem will rely on a federated data collection approach. This approach involves collecting data from multiple sources, including websites, applications, services, DCC and Center teams, and the NIH. In some cases, these metrics and data may be collected automatically through monitoring tools or APIs, while in other cases, manual data collection may be required.

1.2 Common verus Individual Metrics

The CFDE will need to balance the collection of common metrics that apply across the ecosystem with the need for individual metrics that capture the unique contributions and performance of specific components. Common metrics provide a standardized way to evaluate the overall impact and effectiveness of the ecosystem, while individual metrics allow for a more detailed assessment of specific projects, tools, or services.

1.3 Data Privacy and Security

The collection and analysis of metrics within the CFDE will adhere to strict data privacy and security guidelines. Any personal data collected will be anonymized and aggregated to protect individual privacy and confidentiality when used in reporting. Access to sensitive data will be restricted to authorized personnel, and appropriate security measures will be implemented to prevent unauthorized access or disclosure.

1.4 Transparency and Accountability

The CFDE will prioritize transparency and accountability in its metrics and evaluation processes. Data collection methods, metrics definitions, and reporting practices will be documented and made available to stakeholders to ensure clarity and consistency. Regular re-

porting on key performance indicators and progress towards established goals will promote accountability and facilitate data-driven decision-making.

1.5 Continuous Improvement

Metrics and evaluation within the CFDE will be an ongoing process that evolves with the ecosystem. Feedback from stakeholders, user communities, and evaluation results will inform updates to metrics, data collection methods, and reporting practices. Continuous improvement efforts will ensure that the metrics used to evaluate the CFDE remain relevant, meaningful, and aligned with the ecosystem's goals and objectives.

Part I

Metrics

2 Online User Behavior

Understanding online user behavior helps to measure the effectiveness and impact of the Common Fund Data Ecosystem. This section provides example metrics that capture how users interact with the ecosystem's tools, websites, and APIs. By analyzing these metrics, we can gain insights into user acquisition, engagement, and behavior patterns. The metrics are organized into three key areas: User Engagement and Interaction and Behavioral Metrics. Together, these metrics offer a menu that can help teams choose how they quantify user experience and capture how users utilize the ecosystem, highlighting both the successes and potential areas for improvement.

2.1 User Engagement Metrics

2.1.1 Unique Users

Description: The number of distinct individuals who interact with the ecosystem's tools, websites, or APIs over a given period.

This metric provides a basic measure of reach, indicating how many unique users engage with the ecosystem. While it offers insight into the system's broad usage, it does not account for the frequency or depth of interactions. A single user with multiple accounts or devices can lead to inaccurate counts.

2.1.2 New Users

Description: The number of users who interact with the ecosystem for the first time during a specific period.

New users assess the ecosystem's growth and outreach efforts, helping to evaluate the effectiveness of marketing campaigns, user onboarding processes, and user acquisition strategies. However, new users do not guarantee long-term engagement or user satisfaction. Retention and conversion metrics are needed to assess the ecosystem's ability to retain and engage new users.

2.1.3 Returning User Ratio

Description: The proportion of users who return to the ecosystem after their initial visit.

This metric is a strong indicator of user retention and ongoing interest. A high returning user ratio suggests that users find ongoing value in the ecosystem, leading to repeated visits. However, this metric does not reveal the reasons behind the return visits—whether users are coming back for new information or because they failed to complete their tasks during previous sessions. Understanding the motivations for returning is essential for leveraging this metric effectively.

2.1.4 Session Duration

Description: The average length of time a user spends interacting with the ecosystem in a single session.

Session duration serves as an indicator of user engagement, suggesting how compelling or useful the content or tools are. However, longer sessions are not always indicative of positive engagement; they may also signal user difficulties in navigating or understanding the system. The challenge lies in interpreting this metric correctly—understanding whether long sessions reflect sustained interest or user frustration.

2.1.5 Page Views

Description: The total number of pages or screens viewed by users across all sessions.

Page views are a basic measure of activity within the ecosystem, reflecting the volume of user interactions. However, this metric does not provide insight into the quality of those interactions. High page views might indicate robust engagement, but without context, they could also suggest that users are struggling to find the information they need. Interpretation should consider the type of content and user goals. Note that page views can be influenced by factors such as site design, navigation structure, external referrals, and automated processes.

2.1.6 Page Views per Session

Description: The average number of pages or screens a user views during a single session.

This metric helps gauge the depth of interaction with the ecosystem, suggesting how thoroughly users are exploring the available resources. However, like session duration, this metric is subject to misinterpretation. High page views might indicate that users are finding valuable information, but it could also mean that they are struggling to locate what they need. Discerning the underlying cause requires complementary qualitative analysis.

2.1.7 Most Visited Pages

Description: The pages or screens within the ecosystem that receive the highest number of views or interactions.

Identifying the most visited pages helps pinpoint the content or features that are most popular among users. This information can guide decisions about content development, user interface design, and navigation improvements. However, high traffic to specific pages does not guarantee overall success; it may indicate that users are getting stuck or that other areas of the ecosystem are underutilized. Analyzing user behavior on these pages can provide valuable insights into user preferences and needs.

2.1.8 Bounce Rate

Description: The percentage of single-page sessions where users leave the website or tool without further interaction.

Bounce rate indicates the initial appeal or relevance of the ecosystem's landing pages or entry points. A high bounce rate may suggest that users do not find what they are looking for immediately or that the content fails to engage them. However, a low bounce rate does not always signify success; it might be the result of users getting trapped in a complex navigation system. Interpreting this metric requires a careful analysis of user intent and site structure.

2.1.9 Unique Page Views

Description: The number of sessions during which a specific page or screen was viewed at least once.

Unique page views offer a more refined view of user engagement, focusing on the number of distinct sessions in which a particular page or screen was accessed. This metric helps identify the popularity of specific content or features within the ecosystem. However, it does not account for the depth or duration of user interactions. A high number of unique page views may indicate broad interest or effective promotion but does not guarantee user satisfaction or task completion.

2.2 Interaction and Behavior Metrics

2.2.1 Landing Page

Description: The first page or screen a user interacts with when they enter the ecosystem.

The landing page is critical for first impressions, setting the tone for the user's experience. High traffic on a landing page (and there can and usually will be more than one) can indicate effective outreach, but if users frequently bounce from this page, it may signal that expectations are not being met. Optimizing the landing page for clarity and relevance is essential, but the challenge lies in balancing diverse user needs and expectations.

2.2.2 Referer

Description: The URL of the website or tool that directed users to the ecosystem.

The referer metric provides insight into how users discover and access the ecosystem. By tracking the sources of incoming traffic, such as search engines, social media platforms, or partner websites, this metric helps evaluate the effectiveness of marketing strategies and outreach efforts. Analyzing referer data can inform decisions about content promotion, search engine optimization, and partnership development. However, it is essential to consider the quality and relevance of incoming traffic, as well as the potential for referral spam or misleading sources.

2.2.3 Conversion Rate

Description: The percentage of users who complete a specific desired action (e.g., downloading data, signing up for updates) after interacting with the ecosystem.

Conversion rate is a critical measure of the ecosystem's effectiveness in driving users toward specific goals. It reflects the system's ability to move users from exploration to action. However, establishing what constitutes a meaningful conversion can be challenging, and the metric may not fully capture long-term engagement or repeated use. Additionally, conversion rates can be influenced by external factors, such as user expectations or competing resources.

2.2.4 User Logins

Description: The number of times users log in to the ecosystem over a given period.

User logins measure active engagement by tracking how frequently users access their accounts. A higher number of logins suggests ongoing interest and interaction. However, frequent logins without corresponding activity might indicate issues with session management or user difficulties in completing tasks. The challenge is to link logins with meaningful engagement rather than just access.

2.2.5 Total Logged In Users

Description: The total number of distinct users who have logged in during a specified period.

This metric reflects the active user base that is accessing personalized features or tools within the ecosystem. It provides a clearer picture of engaged users compared to general traffic metrics. The challenge is to sustain and grow this base by ensuring that logged-in users find value in their interactions, encouraging ongoing engagement.

2.2.6 New Logged In Users

Description: The number of users who log in for the first time during a specified period.

New logged in users indicate growth in the ecosystem's engaged user base. Tracking this metric helps assess the effectiveness of strategies aimed at converting visitors into active, logged-in users. The challenge is to convert these new users into regular, returning users by providing a compelling and seamless user experience.

2.3 Challenges in Evaluating Online User Behavior Metrics

Evaluating metrics related to online user behavior is inherently challenging due to the variability in user intent, technical factors, and the complexity of interactions. Metrics such as session duration or page views per session can be influenced by multiple factors, making it difficult to draw definitive conclusions. Moreover, these metrics often require contextual interpretation, combining quantitative data with qualitative insights to understand user satisfaction and success.

Another significant challenge is the potential for over-reliance on quantitative metrics, which may lead to an incomplete understanding of user experience. While these metrics can provide a snapshot of user behavior, they do not capture the full spectrum of user motivations, challenges, or outcomes. To address this, metrics should be complemented by user feedback, usability testing, and other qualitative methods that provide deeper insights into user needs and experiences.

Finally, the dynamic nature of online environments adds complexity to metric evaluation. User behavior can change rapidly due to updates in technology, changes in user expectations, or external events. As such, ongoing monitoring and adaptation of metrics are necessary to ensure they remain relevant and accurately reflect user interactions with the ecosystem.

3 Infrastructure Costs

Infrastructure costs refer to the expenses associated with acquiring, deploying, and maintaining the physical and virtual resources required to support an organization's IT operations. These costs can include hardware, software, cloud services, storage, compute, network resources, and other infrastructure components.

Costs should not be viewed in isolation, but rather in the context of the value they provide to the organization. Understanding infrastructure costs is essential for budgeting, resource allocation, cost optimization, and decision-making related to infrastructure investments and upgrades.

In the context of the Common Fund Data Ecosystem (CFDE), monitoring infrastructure costs can help assess the financial sustainability of the ecosystem, identify opportunities for cost savings, and optimize resource utilization to support data sharing, analysis, and collaboration activities.

3.1 Possible Metrics

3.1.1 Hardware Costs

Description: The total cost of purchasing and maintaining physical hardware components, such as servers, storage devices, networking equipment, and peripherals.

Hardware costs are a significant part of the infrastructure budget and include expenses related to acquiring, installing, and maintaining hardware assets. These costs can vary based on the type and quantity of hardware required to support the infrastructure's operations. Monitoring hardware costs helps teams track spending, optimize resource allocation, and plan for future hardware upgrades or expansions.

3.1.2 Software Costs

Description: The total cost of software licenses, subscriptions, and maintenance fees for operating systems, databases, applications, and other software components.

Software costs encompass expenses related to acquiring, renewing, and updating software licenses and subscriptions. These costs can include one-time purchases, annual subscriptions,

and ongoing maintenance fees. Monitoring software costs helps teams manage software expenditures, ensure compliance with licensing agreements, and evaluate the return on investment for software tools and services.

3.1.3 Cloud Costs

Description: The total cost of cloud services, including storage, compute, network resources, and data transfer fees.

Cloud costs are a significant part of the infrastructure budget and include expenses related to using cloud-based services and resources. These costs can vary based on the type and amount of cloud services consumed, as well as the pricing models and billing structures of cloud providers. Monitoring cloud costs helps teams optimize resource usage, control spending, and identify opportunities for cost savings or efficiency improvements.

3.1.4 Storage Costs

Description: The cost of storing data in cloud-based or on-premises storage systems, including object storage, block storage, and file storage.

Storage costs are a key component of the infrastructure budget and can vary based on the amount of data stored, the storage tier or class used, and the data retention policies in place. Monitoring storage costs helps teams manage data growth, optimize storage utilization, and implement cost-effective storage solutions.

Note that storage costs can be further broken down into costs associated with different storage classes or tiers, such as standard storage, infrequent access storage, and archival storage.

3.1.5 Compute Costs

Description: The cost of using compute resources, such as virtual machines, containers, and serverless functions, to process data and run applications.

Compute costs are a significant part of the infrastructure budget and can vary based on the type, size, and duration of compute resources used. Monitoring compute costs helps teams optimize resource allocation, scale computing capacity based on demand, and control costs associated with running applications and workloads.

3.1.6 Network Costs

Description: The cost of network resources, such as bandwidth, data transfer, and network egress fees, required to transmit data between components of the infrastructure.

Network costs are an essential part of the infrastructure budget and can vary based on the volume and frequency of data transfers, as well as the network performance requirements of the system. Monitoring network costs helps teams optimize data transfer efficiency, reduce latency, and manage expenses associated with network usage.

Note that network costs can be further broken down into costs associated with different resources, datasets, or services.

4 Website and Application Technical Performance

The CFDE teams spend significant time and effort developing and maintaining websites and applications that comprise the CFDE ecosystem. These websites and applications serve as the primary interfaces for users to access data, tools, and resources. Providing metrics that evaluate the technical performance of these digital assets recognizes those efforts and helps identify areas for improvement.

4.1 Metrics

4.1.1 Page Load Time

Description: The time it takes for a web page or application to load and display content to the user.

Page load time is a critical metric for user experience, as it directly impacts user engagement and satisfaction. Slow-loading pages can lead to increased bounce rates and decreased user retention. Optimizing page load times involves various strategies, such as reducing image sizes, minimizing server requests, and leveraging caching mechanisms. Monitoring page load times across different devices and network conditions helps ensure a consistent and responsive user experience.

4.1.2 Uptime

Description: The percentage of time a website or application is operational and accessible to users.

Downtime can significantly impact user trust and satisfaction, especially for critical services or resources. Uptime metrics provide insights into the reliability and availability of digital assets, helping teams identify potential issues and prioritize maintenance tasks. Monitoring uptime over time can reveal patterns of instability or performance degradation, guiding efforts to enhance system resilience and minimize service disruptions.

4.1.3 Server Error Rate

Description: The frequency of errors generated on the server side in response to user requests.

Server errors, such as HTTP 500 responses or database connection failures, can disrupt user interactions and degrade system performance. Monitoring server error rates helps teams detect underlying issues in the back-end infrastructure, enabling timely troubleshooting and resolution. Addressing server errors promptly is essential for maintaining service quality and preventing negative user experiences.

4.1.4 Client Error Rate

Description: The frequency of errors generated on the client side during user interactions with a website or application.

Client errors, such as JavaScript exceptions or rendering failures, can degrade the user experience and hinder interaction with digital assets. Monitoring client error rates helps teams identify problematic areas in the front-end codebase, enabling targeted debugging and optimization efforts. Addressing client errors promptly can enhance user satisfaction and improve overall system performance.

4.1.5 Completed Jobs

Description: The number of successfully completed tasks or processes initiated by users within a web application.

Completed jobs reflect user engagement and interaction with the functionalities provided by a web application. Tracking completed jobs helps teams understand user behavior, usage patterns, and feature adoption rates. Analyzing completed jobs data can inform decisions about feature prioritization, usability improvements, and performance optimizations to enhance the overall user experience.

Part II

DCC Metrics

Part III

Center Metrics