Managing Large-scale Cancer Research Data Programs

Dina Mikdadi, MPA[2], Chelsea Owens, MS[2], Juergen Klenk, PhD[2], Bhavani Singh, MS[2], Eric Barner, BS[2], Angela Maggio, MS[2], Mike Warfe, MBA[3], Tanja Davison, PhD[1], Ina Felau, MS[1], Erika Kim, PhD[1]

Published Abstract Number: 3568

[1]National Cancer Institute [2]Deloitte Consulting LLP [3]Frederick National Laboratory

Introduction and Background

The volume of data produced by the cancer research community continues to grow rapidly. Efficient data collection, curation, and sharing are expected to enable researchers to synthesize larger datasets and apply novel methods to discover new patterns for diagnosis, treatment, and care of disease. Consequently, data platforms that support these functions are an important because they serve to democratize and increase the utilization of biomedical data and analytic tools to accelerate cancer research outcomes. The Cancer Research Data Commons (CRDC) is the Cancer Moonshot program and NCI's premier investment in such a platform, and it is one of several NIH cloud data platforms designed to provide secure access to an expanding collection of curated biomedical and clinical research data.

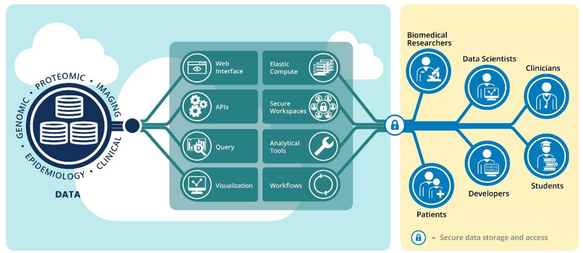

Fig. 1. NCI's CRDC Ecosystem

Due to the complexity and scale of this ecosystem (Fig. 1), CRDC requires a carefully designed data governance framework to enable CRDC's goals to sustain a collaborative data environment where researchers can access timely, accurate, and relevant cancer data.

CRDC Data Governance Goals

1) Establish Clear Roles and Responsibilities
2) Design and Distribute Effective and practical Policies, Standards, and Procedures (PSP)
3) Define Measurable Performance Outcomes
4) Enable Transparent Communication Across Stakeholder Groups

Participation Incentives

1) Encourages shared understanding of data standards decisions across component teams
2) Brings individual component teams' own best practices to the table
3) Provides opportunities to influence data management architectures and common data curation workflows

Acknowledgments

The CRDC Sustainability Implementation team would like to thank the NCI CBIIT leadership including Dr. Tony Kerlavage, Dr. Jill Barnholtz-Sloan, Mr. Jeff Shilling, Dr. Jaime Guidry Auvil for their guidance, Dr. Tanja Davidsen, Dr. Erika Kim, and Ina Felau for their commitment as government leads for this study, and all CRDC staff and contractors for their assistance in our research and data collection phases of this work that enhanced our understanding of the complex CRDC ecosystem.

Designing a CRDC Data Governance Board

Environmental Scan and Gap Analysis: The team conducted a series of component interviews and process reviews to understand existing data management and governance & identify critical gaps. We reviewed CRDC process data and documentation for 11 component teams.

Governance Framework Design: We considered many designs for the CRDC Data Governance Framework, including fully centralizing or decentralizing decision-making bodies (Fig. 2). Implementing a federated Data Governance framework ensures broad perspective inclusion and consistency across a complex data ecosystem.

Launch: The CRDC Data Governance Board officially launched in August 2023. Currently established CRDC sub-groups include the Enterprise Architecture Review Team, Submission Review Pilot Team, and Data Standards Services Standing Committee. The Data Governance Board's initial focus areas for 2023-24 are represented in Fig. 3. There will be a strong continued focus over the next year on kicking off the remaining governing sub-groups and documenting broad policies, standards, and procedures that enable long-term sustainability and scalability of the CRDC.

	Centralized Data Governance	Federated Data Governance	Decentralized Data Governance
Description	• Data Owner and Data Stewards are full-time, permanent roles, managed within an organization • All data maintenance is performed within the centralized team	• Data Governance Board has responsibility for policy and governance, including a smaller dedicated team to oversee the data governance program • Data maintenance is primarily performed in the local functional areas but also can occur at the enterprise level	• Sub-organizations fully govern their data • Centralized governance committee acts as a sounding board for enterprise-level issues; no permanent resources • Data maintenance is performed exclusively in the local functional areas
Benefits	• Strongest focus on policy setting and driving enterprise-level decisions	• Places the Data Owners and Data Stewards in charge of their data, keeping ownership as close to the source as possible • Allows for centralized consideration of policies, standards, and procedures that affect multiple data teams	• Data issues can be resolved on an application-by-application basis • Less resource intensive
Challenges	• Difficult to implement successfully in a highly decentralized organization • The central information management function may not fully understand the data requirements and users' needs • Most resource intensive	• Requires influence and negotiation skills to manage setting enterprise-level decisions and ensuring commitment / adoption to new policies, standards, and procedures	• No central head to drive consistency and enforce compliance or commitment • Potential loss of focus/energy

Fig. 2. Centralized, Federated, and Decentralized Governance Frameworks (above)

Data Governance Board (DGB)	Component Maturity Assessment	Standard Processes & Metrics	User Engagement	Data Lifecycle Optimization Support
• DGB Charter, kick-off, facilitation & coordination • Convene and support initial Standing Committees and Working Groups	• Conduct Component Maturity Assessment across multiple key functional domain areas	• Design common SOP Template and draft priority standard operating procedures • Identify common financial metrics & reporting process	• Identify current / unreached user communities and develop engagement strategy for new feature development	• Support development of Business Process Optimizations across six Data Lifecycle Stages

Fig. 3. 2023-24 CRDC Key Data Governance Focus Areas (above)

CRDC Data Governance Board (DGB) Structure

In 2023, CRDC chartered a Data Governance Framework (Fig. 4) of Committees and Working Groups across all CRDC components to: Enable diverse data type sharing; provide secure data access; optimize common infrastructure components and functions; and adhere to FAIR data principles (Findable, Accessible, Interoperable, and Reusable).

1) The top Leadership tier includes existing National Cancer Institute Leaders. CBIIT and NCI Leadership are important stakeholders in ensuring that CRDC governing decisions align with broader strategy, policy, resource allocation, and guidelines.

2) CRDC Leadership and decision makers consist of existing CRDC leadership who provides overall responsibility, prioritization, and oversight of the DGB

3) The CRDC Data Governance Board includes a Chair and Co-Chair, Standing committees, rotating Functional Working Groups, and Rotating Governance Roadmap Prioritization SMEs. The board is responsible for operations, strategy, and guidance of overall CRDC governing decisions and functions.

4) Component-specific Local Governance teams provide updates on ongoing governance activities, processes, and successes to inform CRDC's community and leadership.

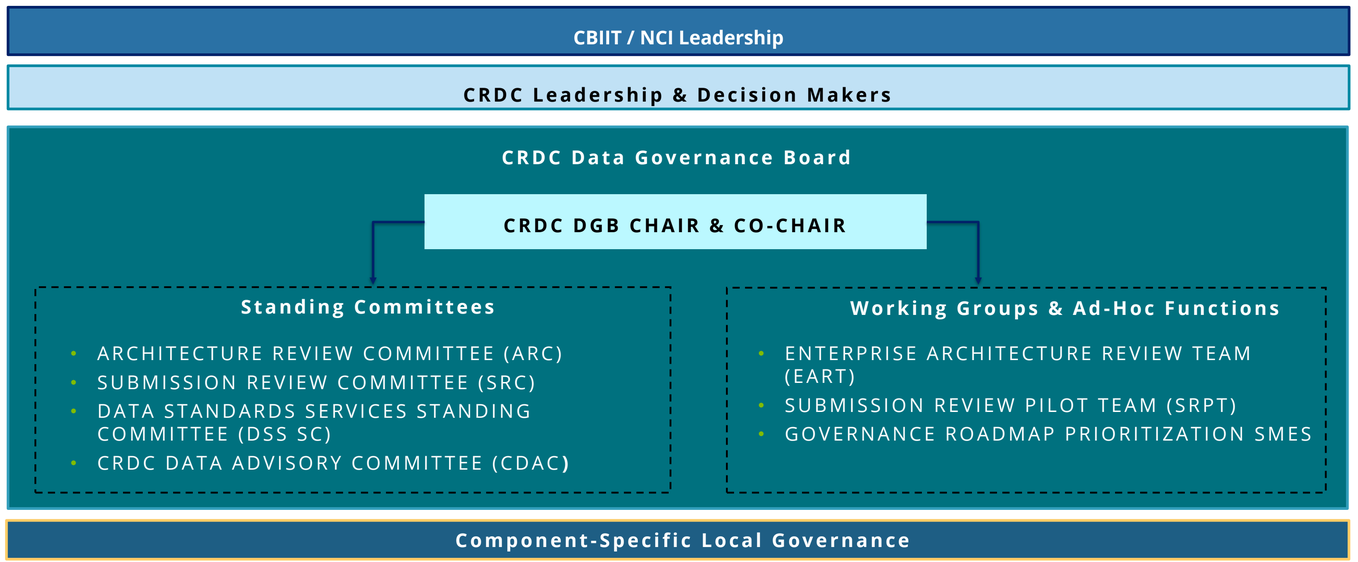

Fig. 4. CRDC Data Governance Board (DGB) Org Structure