CFDE November Meeting

Website: https://nih-cfde.github.io/2022-nov-meeting/
Slides:
https://docs.google.com/presentation/d/11bWhk-bo7t4UDT8U40DdvPOJG2TdtEJXEa1rCQc48Kg/edit#slide=id.p

November 9, 2022, 1:30pm ET

Open Platform Discussion

Session Chair: Aleks Milosavljevic, Adam Resnick
Reporter: Eric Wenger

Meeting Summary

Overview: To review and discuss an ecosystem approach to accelerate and/or empower discovery and translation using starting prompts.
Goal: The starting prompts that began discussions included FAIR knowledge to cut across silos and facilitate discovery, beyond use cases, workflows and analytics, define a governance and tools framework for data exchange and access, and grow the CFDE knowledge base and build the CFDE brand.
Outcomes: We discussed FAIR data, information, and computable knowledge to be communicated as knowledge graphs, and utilizing a Knowledge Graph working group where development of an operational framework for knowledge graphs can occur. We can focus on identifying model studies from published papers instead of use cases. The playbook is a step toward having a unified CFDE approach for workflow and analytics. Many DCCs have controlled access data requiring special permissions required for data access, but the CFDE should define a framework to address the end-to-end data access needs across the broad community of DCCs and users. Best practices for growing the CFDE knowledge base and building the CFDE brand include developing a public library of RFCs, SOPs, best practices, and publications, leveraging existing communication channels/communities, and creating a public CFDE Q&A site.

Notes

Introduction

- Review of an ecosystem approach should accelerate and/or empower discovery and translation.
- Start high-level prompts for discussion.
- We have opportunities for DCCs collectively to contribute to a unified ecosystem approach that drives discovery and value.
- The gist of the starting prompts is whether knowledge can be regarded as first class artifacts in the same way as data. If so, this requires a new scope for the CFDE focus.

Detailed review of starting prompts

FAIR knowledge to cut across silos and facilitate discovery: FAIR data but also FAIR information and computable knowledge (may be communicated as knowledge graphs). Discussion and feedback below:

- Agreed in the description of knowledge as computable knowledge.
- Knowledge under discussion is generalizable and supported by assertions.
- There is a need for an operational framework for CFDE knowledge graphs.
    - The Knowledge Graph working group is the forum (which is in the process of maturing) where development of an operational framework for knowledge graphs can occur. This WG is where alignment needs to occur on the roadmap (development, however, would be taken on separately by other DCCs).
    - How can a knowledge graph in development in the CFDE empower, for example, Kids First? This ties into a separate comment that a knowledge graph implementation needs to meet the scientific needs of investigators.
    - It is important to adequately document and communicate for the public the details about the knowledge graph.
    - NIH has approved the creation of a CFDE Knowledge Core.

Beyond use cases: Focus on identifying model studies from published papers instead (make it faster, better, empower wider research community to conduct specific types of research). Discussion and feedback below:

- Goal is that CFDE resources can be demonstratively used to reproduce research workflows in a way that is quicker and easier to execute by an individual or comparatively smaller group of investigators. Papers are “fixed in time,” but CFDE has the opportunity to make these dynamic.
- We should explore new opportunities and leverage existing research.
- This is outcome-focused; not necessarily an elimination of use cases.
- An opportunity exists to do integration testing (helps to identify gaps in the existing workflow/guides infrastructure development).
- Consider a link to outreach and training since insights from integration testing can inform others. The community still does not know where to go to CFDE for what resources. Each DCC has a user community, e.g., the Kids First community had research and

publication needs that drove their use cases.

Workflows and analytics: How can we better support the “use” of DCC datasets for downstream analyses in the context of the CFDE’s cloud resources? Discussion and feedback below:

- Each DCC is creating its own framework for analysis and analytics. There is a challenge in having a unified CFDE approach. The playbook is a step towards resolving this, which includes FAIRness of API access and accessible workflow in the SigCom application.
- Workflows enable multiple categories of use cases.
- Do I need to bring all of my data to your workspace to integrate our data for analysis? It is important to leverage existing standards (and build them out further if incomplete) and work towards “operational” interoperability. We should not plan on just building bespoke CFDE solutions.
- Self-serving current CFDE workflows are not ideal; better to reproduce workflows represented in publications. The challenge is that labs are often averse to simply taking off-the-shelf products since they want to customize.

Define a governance and tools framework for data exchange and access: Many DCCs in CFDE have a combination of controlled access human subject research data and open access data sets and resources. Can we define and implement emerging standards and resources (RAS, DRS, GA4GH)? Discussion and feedback below:

- Many DCCs have controlled access data requiring special permissions required for data access.
- CFDE should be the testing ground for a holistic framework that addresses the end-to-end data access needs of investigators. Address the needs across the broad community of DCCs and users, not just addressing in a DCC-specific bespoke manner.
- Consider how users can access data, how users can analyze the data (avoiding downloading all files), how to address the challenge of data streaming across different cloud platforms (egress costs, including across different regions), and provide a roadmap for users to follow in clearly understanding how to handle this.
- Does the DRS/RAS framework work at scale? Is it essentially just an authorization/download mechanism? The new DCCs said users are not comfortable with compute-in-place since they have concerns about transparency in cost (or projected cost), availability of data in the cloud platform, and whether or not the steps are clear and understandable. There are many policy and usage questions that are currently unresolved and need to be addressed.

Grow the CFDE knowledge base and build the CFDE brand: Agree on best practices and standards within CFDE (RFC process, CFDE RFC publications) and then propagate to new DCCs and wider communities. Discussion and feedback below:

- Develop a library of RFCs made available to the community.
- CFDE has the opportunity to emphasize and publicize RFCs, SOPs, best practices, and publications and elevating their importance, further empowering the transfer of knowledge and potentially “build the CFDE brand.” For example: Stack Exchange for CFDE.

- Currently unpublished RFCs could be released as publications.
- Leverage existing communication channels/communities and incorporate CFDE or create new communication channel(s).
- Branding is important and there is a need to create CFDE-specific resources, such as a website or public Q&A site (need to be mindful of the challenge of using some existing platforms where users could be tagged negatively for a question or a response). The existing community and its own feedback loop can organically drive growth.
- Another opportunity is to have a CFDE bridge between ICs.
- CFDE has the opportunity to specify the RFAs for new Common Fund programs. For example, the Bridge2AI program has the potential to create standards for future data use.
- Each DCC has its own individual community. There is a much larger group of users who are individually using the data, so we have the opportunity to be strategic about bringing our users together. For example, Kids First invited INCLUDE to attend ASHG due to complimentary interests.