SIMB | Society for Industrial Microbiology and Biotechnology

Accelerating the Bioeconomy Through a **Pre-Competitive Knowledgebase** for Biomanufacturing and Biotechnology

Supported by Schmidt Sciences

SIMB 2024. Accelerating the Bioeconomy Through a Pre-Competitive Knowledgebase for Biomanufacturing and Biotechnology. Society for Industrial Microbiology and Biotechnology. doi:

EXECUTIVE SUMMARY

Biotechnology and biomanufacturing are poised to revolutionize a circular bioeconomy that leverages biological systems to produce products across multiple sectors and address global challenges such as climate change, food innovation and supply chain resilience. This will boost global economic competitiveness, and provide new employment opportunities. Yet, to build and grow a sustainable bioeconomy, foundational capabilities are urgently needed including being able to 1) engineer biology in a more facile and accelerated manner, 2) develop, pilot and scale biomanufacturing, 3) facilitate data access and sharing, and 4) fully access the global biotechnological knowledge, talent, and services.

The need for increased data sharing for biomanufacturing and biotechnology has been long recognized as a means for accelerating the bioeconomy through providing technical, operational and supportive information to advance product development and commercialization, avoid pitfalls and guide workforce development. Much of this information currently remains in inaccessible (i.e. company) data repositories and in unstructured forms that limit its use for e.g. Artificial Intelligence applications. As recommended in the 2017 Council on Competitiveness report, "develop widespread and easily accessible knowledge bases of principles, methods, processes, successes and failures to more quickly deliver helpful information to stakeholders. Industry access to central scientific and technical resources will help experts develop and deliver new, innovative products to the market. This will improve the maturation and impact metrics of the bioeconomy and assist in the technology innovation pipeline from development in the laboratory to scaling-up in the manufacturing plants on to consumer outlets."[1] More recently, the 2022 White House Executive Order on advancing biotechnology and biomanufacturing called for the establishment of "a Data for the Bioeconomy Initiative (Data Initiative) that will ensure that high-quality, wide-ranging, easily accessible, and secure biological data sets can drive breakthroughs for the United States bioeconomy."[2] If such a data system can be established, it is expected that bio-based product development and commercialization will be accelerated and the bioeconomy will grow significantly.

The Society for Industrial Microbiology and Biotechnology (SIMB), with the support of Schmidt Sciences, convened three workshops in 2023 and 2024 that brought together thought leaders and experts from industry, academia, government and non-profit organizations to discuss how a Pre-Competitive Knowledgebase (PCK) for biotechnology and biomanufacturing could facilitate data access and sharing that would accelerate and lower the cost for the development of bio-based products, towards an ultimate goal of a 50% reduction. This document serves as a report out from those workshops to inform on what the establishment of a PCK would entail, its value and our recommendations for next steps.

The PCK is envisaged to be a structured repository and facilitation platform that addresses the challenges of making pre- and post-competitive data available, which has otherwise been sequestered or largely ignored. This includes data and models on organisms, bioreactors, downstream processing methods, and regulatory impacts.[3] By focusing initially on technical data types and models related to fermentation and bioprocessing, and establishing standards for data quality and metadata, the PCK aims to provide reliable, curated information to support research and innovation.

1 *https://competeorg.wpengine.com/wp-content/uploads/historical-program-reports/emcp-leverage-bioscience-2017.pdf*

2 *https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/*

3 *https://www.whitehouse.gov/wp-content/uploads/2023/12/FINAL-Data-for-the-Bioeconomy-Initiative-Report.pdf*

The PCK should bridge foundational research and application. Through standardizing terminology, fostering community dialogue, and serving as an educational tool, the PCK will enhance collaboration, streamline regulatory compliance, and address challenges across different Bioindustrial Manufacturing Readiness Levels. To achieve this, mechanisms for rewarding academia-government-industry partnerships, student engagement in data curation, and industry participation need to be established, while addressing challenges such as funding sustainability, incentivization, cyber-security and data integration across diverse sources and disciplines.

The PCK should be structured to meet the needs of the biomanufacturing and biotechnology research and development community by incorporating clear organizational infrastructure, digital object management, and technology mechanisms. The PCK should emphasize the importance of roles and permissions, data curation, metadata schema, and provenance information to ensure data quality and transparency, while also facilitating ease of access, data sharing agreements, and tracking mechanisms to monitor usage metrics and impact. Active participation and tangible benefits must be realized, necessitating early identification of successes and the development of case studies showcasing the value of the PCK. Incentives for participation, such as financial rewards, credit schemes, or demonstrable benefits like forming partnerships or gaining access to new technologies, should be implemented to engage both individual contributors and organizations, with mechanisms in place to ensure fair distribution of costs and benefits. Measures for tracking impact and effectiveness, including quantitative metrics, testimonials, user surveys, and documentation of organizational diversity, should be implemented to assess the growth and value of the PCK over time. Additionally, its success also relies on its ownership, governance, and management. Funding can come from various sources, and the steering team will be responsible for design, management, and long-term funding, ensuring data quality, version control, while safeguarding cybersecurity and commercial interests. Workshops that can generate testimonies from users will be essential to highlight the PCK's value, along with mechanisms in place to track usage securely and maintain data integrity.

The PCK would provide a unique platform for the sharing and use of pre- and post-competitive data that has otherwise been inaccessible to those researchers and companies seeking to commercialize bio-based products and thus has an important role to play in accelerating the growth of bioeconomy, and thus needs to be established as soon as is possible. The development of the PCK is projected to span four years beginning with stakeholder research and design in Year 0, followed by infrastructure building and data population in subsequent years, aiming for deployment with at least 10 organizations actively using the platform by Year 3. The initial focus will be on establishing a minimum viable product (MVP) version of the PCK, demonstrating early successes to secure sustained funding and planning for long-term sustainability post-roadmap. Beyond the initial roadmap, the PCK aims to evolve into a comprehensive resource encompassing regulatory, supply chain, safety, and biomedical data, fostering broader adoption and increased partnerships to enhance its value to the bioeconomy.

The PCK should either be folded into existing data platform efforts or be funded independently and funds should be made available in a short time frame for scoping and stakeholder interviews to refine the vision and MVP ideas put forth here prior to funding the development work.

The expedited timeline and funding suggested for the PCK (~$100M USD over the next 4 years) was based on careful consideration of international investments in infrastructure and ecosystem to collect high-quality bioprocess scale-up data. This timely investment would enable the development of domain-specific large models on process development and scale-up. Many innovators in the US - from startups and academic environments - will gravitate towards using such models for most of their scale-up needs. Higher usage will make these models more robust and accurate, and in turn more preferred, creating a virtuous cycle that will require an inordinate amount of effort and funding to surmount. The risks associated with not having a US based PCK that can inform large models include the "offshoring" of not just data but also ideas from the growing multi-trillion US bioeconomy. The proposed PCK is a critical first step essential in the next few years for the US to remain competitive in bioprocess scale-up and manufacturing and thereby conserve and grow the US bioeconomy.

INTRODUCTION

Biotechnology and biomanufacturing are poised to revolutionize the circular bioeconomy. Leveraging biological systems to produce goods and services at a commercial scale holds immense potential to revolutionize industries including plastics, fuels, medicines, and food, contributing to sustainable alternatives and addressing challenges such as climate change, food innovation, and supply chain resilience. The economic impact of the bioeconomy is substantial, with the National Academies of Sciences, Engineering, and Medicine (NASEM) estimating a $959 billion total economic impact in 2016. A 2020 report from McKinsey suggests the global bioeconomy could generate $2 trillion to $4 trillion in annual economic impacts by 2030-2040. Yet in order to build and grow a sustainable bioeconomy, foundational capabilities are urgently needed such as being able to 1) engineer biology in a more facile and accelerated manner, 2) develop, pilot and scale biomanufacturing, 3) facilitate data access and sharing, and 4) fully access the global biotechnological knowledge, talent, and services. Additionally, three other critical gaps were recognized by the President's Council of Advisors on Science and Technology[4]: insufficient manufacturing capacity, regulatory uncertainty, and an outdated national strategy.

Today, we have already seen significant investments being made to advance engineering biology through for example, the Agile Biofoundry[5] and BioMADE for new pilot and scaling fermentation facilities[6]. These are significant strides forward, but are insufficient to unlock the full potential of the bioeconomy without additional investments in foundational capabilities. The Biden-Harris administration has placed a significant emphasis on building the bioeconomy workforce of the future, and has highlighted the importance of education and training in biotechnology and biomanufacturing in technician education and degree programs, including cross-cutting expertise such as computer systems analysis, software development, data science, and bioinformatics.[7] BioMADE is pairing industry-driven competencies with program development through K-12 schools, community colleges, universities, and professional development organizations[8], and NIST Funding will support a NIMBL pilot program with Merck, Pfizer, Pitt Community College, and the North Carolina Biotechnology Center.[9]

One of the foundational capabilities that has received relatively less attention and funding than the others highlighted above is data sharing and

4 *https://www.whitehouse.gov/wp-content/uploads/2022/12/PCAST_Bio-manufacturing-Report_Dec2022.pdf*
5 *https://agilebiofoundry.org*
6 *https://www.biomade.org*
7 *https://www.whitehouse.gov/wp-content/uploads/2023/06/Building-the-Bioworkforce-of-the-Future.pdf*
8 *https://www.biomade.org/education-workforce-development*
9 *https://www.niimbl.org/workforce*

access. As summarized in the Schmidt Futures report on "The U.S. Bioeconomy: Charting a Course for a Resilient and Competitive Future"[10] opportunities exist for establishing a shared and accessible data infrastructure to propel collaborative research and development to accelerate industrial biotechnology. Currently, the biomanufacturing and biotechnology data landscape is highly fragmented with data in many formats, qualities and repositories across many institutions, both private and public. There is little, if any, interconnectivity between these data resources. Given their experience in scale-up and commercialization, industry can play a key role in sharing knowledge and expertise through public-private partnerships with government and academia. The sharing of precompetitive data and software tools for bioengineering will reduce the development costs (currently estimated at >$100 million) and time (currently estimated at 10 years) for bio-based products thereby accelerating the scale-up and commercialization of biotechnological processes.

Bringing together and integrating diverse data types from biomanufacturing and biotechnology including omics measurements, process measurements (both bioreactor and downstream operations), analytics, imaging, and techno-economic and lifecycle assessments, along with process management, standardized terminology, and curation will drive the development of models that can be used to reduce time and resources to scale-up.[11] Mechanisms and governance are required to do this in a manner that aligns with the FAIR (findable, accessible, interoperable and reusable) data principles and makes data useful to all, secure, and trustworthy.[12] Additionally, there needs to be a standardized anonymization policy and defined practice on how to productively use models built on the data without revealing the primary data itself. Such a system will enable those working in the bioeconomy to search and find data to guide and accelerate their work in scaling-up bio-based processes.

The Society for Industrial Microbiology and Biotechnology (SIMB), with the support of Schmidt Sciences, brought together thought leaders and experts from industry, academia, government and non-profit organizations to discuss what a Pre-Competitive Knowledgebase (PCK) for biotechnology and biomanufacturing could deliver to meet needs, what data should be included, how data would be identified and inputted, curation and standards, how different participants could be incentivized and how intellectual property and partnerships could be managed. Three workshops were held in 2023 and 2024.

A PCK for biotechnology can play a key role accelerating the bioeconomy through contributing to the acquisition, use, sharing, and analysis of data for the bioeconomy in private, public, academic, industrial, and federal contexts to promote best approaches and avoid repetition and pitfalls. Through the development of community-driven standards, structured data and access mechanisms, the PCK can enable interoperability and integration through private-public partnerships to increase data sharing among those working in and for the bioeconomy. The PCK can foster the aggregation, analysis and synthesis of existing and future pre-competitive data. It would also be valuable to develop new tools to analyze and visualize this data allowing researchers to extract the most information.

While the advantages of developing and deploying a PCK platform are abundant and clear, many challenges exist in the practical implementation, management and upkeep of the system. Biological

10 *https://www.schmidtfutures.org/wp-content/uploads/2023/05/Bioeconomy-Task-Force-Strategy-4.14.22-4.pdf*

* From 2024, the bioeconomy work formerly housed in Schmidt Futures continues in Schmidt Sciences, a new charitable organization that evolved from the core science work achieved at Schmidt Futures over the past five years.

11 *https://www.whitehouse.gov/wp-content/uploads/2023/12/FINAL-Data-for-the-Bioeconomy-Initiative-Report.pdf*

12 *https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf*

data is inherently extremely broad in nature leading to different types of data formats and structures. Furthermore, data terminology lacks standardization across the industry. There are also concerns around data security, intellectual property and incentivization of participation - just to name a few. This report outlines the key considerations, challenges and recommendations that were discussed in the series of workshops in establishing and operating a PCK for biomanufacturing and biotechnology.

WORKSHOPS AND EMERGING THEMES

Three workshops were held that were organized by SIMB in 2023 and 2024. The first workshop was held at the Hyatt Regency Minneapolis on Saturday, July 29, 2023, and focused on what a biotechnology knowledgebase that advances biotechnology and biomanufacturing innovation could comprise of, what data is most impactful to include, how to incentivize data providers, how data will be accessed (user interface), and how data security will be managed. The second workshop was held at the Naples Grand Beach Resort, Florida on Friday October 27, 2023, and focused on how public-private partnerships for the acceleration of translation of basic research to application can be strengthened, mechanisms of establishing partnerships, open data sharing and reporting, streamlining tech transfer and contracting, and managing intellectual property. A third, smaller working group was convened on Monday, January 22, 2024, at dsm-firmenich in Columbia, MD to review the findings, add additional information to close gaps, refine the draft report and to ensure that the challenges and recommendations are articulated. A list of participants and the workshop agendas are to be found in Appendix 1 and Appendix 2, respectively.

Initial Sentiments and Perceptions

At the outset of the workshop series, key questions were raised as to what the value of a PCK is, what is pre-competitive knowledge, what data should a PCK house, how would it be used, what would the safeguards in terms of security and data integrity look like, and how would the PCK be managed, funded and sustained as the core themes for exploration during the workshops.

Why is a PCK valuable and what would it enable?

A PCK would enable broader use of existing data to accelerate time-to-market and decrease development costs for products produced using biotechnology and biomanufacturing. Growth of the bioeconomy will be realized through technological advancements and successes in biomanufacturing. Currently though, time to market and costs for developing bio-based products are at least 10 years and >$100 million[13]. There are already various approaches being pursued to accelerate this development and reduce costs e.g. BioMADE, Agile BioFoundry, Advanced Biofuels and Bioproducts Process Development Unit, entities that provide scale-up capabilities to move processes from lab to (pre)-commercialization. Yet there still exists much information across the biomanufacturing and biotechnology research and development enterprise that has not been used as widely as it could be and which also could serve to reduce barriers to bio-based product development. This is due to that information being largely sequestered away behind firewalls (e.g. in company data stores) or in hard to access formats (e.g. lab notebooks) that prevent broader use of that data. Having a data repository and data facilitation platform in

13 Nielsen, J. and Keasling, J. D. 2016. Engineering Cellular Metabolism. *Cell* 164: 1185-1197. *https://doi.org/10.1016/j.cell.2016.02.004.*

the form of a PCK would enable the structuring, searchability and use of that data to propel the bioeconomy.

The access and use of precompetitive data can reduce time to solve technical challenges, process development, provide the data required for technological advancements and reduce redundancy. Having multiple data sets from existing and historical bio-based products will allow horizontal and vertical analyses in new ways that can reveal new information. The provision of this data, and the use mechanisms, can also foster new public-private partnerships that distribute expertise. Furthermore, this can also drive workforce development in training for how to generate, use and disseminate this data in more open and/or collaborative means.

The ultimate goal of the PCK would be to significantly contribute to the reduction of the time and cost for developing bio-based products, to drive towards an overall 50% reduction, that is, from 10 years and $100 million to 5 years and $50 million.

What is pre-competitive data?

Pre-competitive data is defined here as research findings, scientific information, models or datasets that encompasses foundational knowledge, experimental data and technical insights that are essential for advancing scientific understanding and innovation, with the aim of benefiting the broader scientific community and accelerating progress towards common goals without conferring competitive advantages to any single entity. We consider proprietary data to be data which is restricted or protected via trade secrets, institutional knowledge, or other mechanisms that do not result in public release. Pre-competitive data can be both published and unpublished data and there is a lot of data that exists that is not findable or searchable to which the PCK can add value. Published data includes that which has been released in journals, patents, reports, presentations, white papers, application notes, methods, regulatory filings, market intelligence reports, dissertations/theses, government publications, grant applications, industry surveys and data in databases. Unpublished data or data that is not readily findable includes negative results i.e. the results of experiments, investigations or analyses that do not yield data that is further actionable on or that gave failures, market data and trends, consumer sentiments, supply chain information, regulatory path information, industrial scale operations or recipes, equipment designs and waste management. Often these datasets are too small to paint the full picture and therefore not valued enough for peer-reviewed publications. However, these small datasets can be impactful and save substantial funding, and even more importantly, time by ensuring that the same mistakes are not repeated.

Specific data classes that should be considered pre-competitive data for the PCK include:

- Information on organisms, their cultivation conditions and physiology, metabolites they produce or consume
- Omics data (genomics, transcriptomics, proteomics, metabolomics)
- Bioreactor specifications and operational modes
- Process operations, feeding regimens, control set points
- Downstream processing methods, recovery yields
- Other equipment specifications, settings, operations
- First-principle and empirical models of bioprocesses and specific unit operations

- Raw material specifications and pricing
- Safety, containment and disposal of by-products
- Regulatory impacts
- Environmental impacts

There is a wealth of precompetitive data that exists across public and private repositories. While some of this data is already accessible (e.g. genomics data at NCBI), much is inaccessible or is only accessible through hard-to-find means. The majority of the data does not include sufficient metadata, is not structured or lacks details that make the interoperability and reuse of this data very limited without additional curation. The goal of the PCK would be to enable access to this data in a format that can be used to stimulate advancement and cooperation to accelerate the bioeconomy to decrease costs and time for bio-based product development.

There is also data that can be described as post-competitive data, namely that data which has already been used and generated from commercial deployment. This is also very valuable data to include in the PCK to again enable researchers and product developers to find productive paths to commercialization and eliminate those that have already been pursued (and are tied up in IP) or which have failed. Key examples of post-competitive data include the development, scale-up and commercialization data for historical products that are no longer on the market.

With any of the data in the PCK, trustworthiness and integrity of the data is paramount in the value of this data and usability by those that will use the PCK. The establishment of robust standards, quality control of the data and if available, including models along with the data, can help alleviate this issue and instill confidence in the data. Developing and making available comprehensive, structured, uniform datasets can drive the training of models for Artificial Intelligence/Machine Learning that can generate new insights.

Challenges in Establishing and Sustaining the PCK

The success and value of the PCK will only be as good as the quality and usefulness of the data that is contained within it, who provides the data and in what format, how this data can be accessed and used, security of the data and how the PCK is managed, funded and sustained. There will need to be incentives for data provision and whilst initial data will be provisioned from public sources, others will come from private repositories and reports. What data is most valuable to include and what is eliminated? Companies may be unwilling to share their data or the metadata that makes the data useful. Maintaining motivation for data providers and users will be key, and it will be important to identify and communicate early success stories. Verifying the quality of data will be challenging and data curation will be needed. It is unclear as to who the main user groups of the PCK will be and will the mechanisms of the PCK align with their needs. It is also difficult to define how the PCK will be governed, funded and sustained through different phases of development and deployment. These issues and challenges are elaborated on in the next sections.

Theme 1: Accelerating the Translation of Basic and Applied Research

The growth of the bioeconomy is dependent on a steady stream of products produced using biomanufacturing and biotechnology entering the marketplace and their increased usage. This requires a seamless and streamlined translation of basic research towards applications and the bringing together of data, technologies and

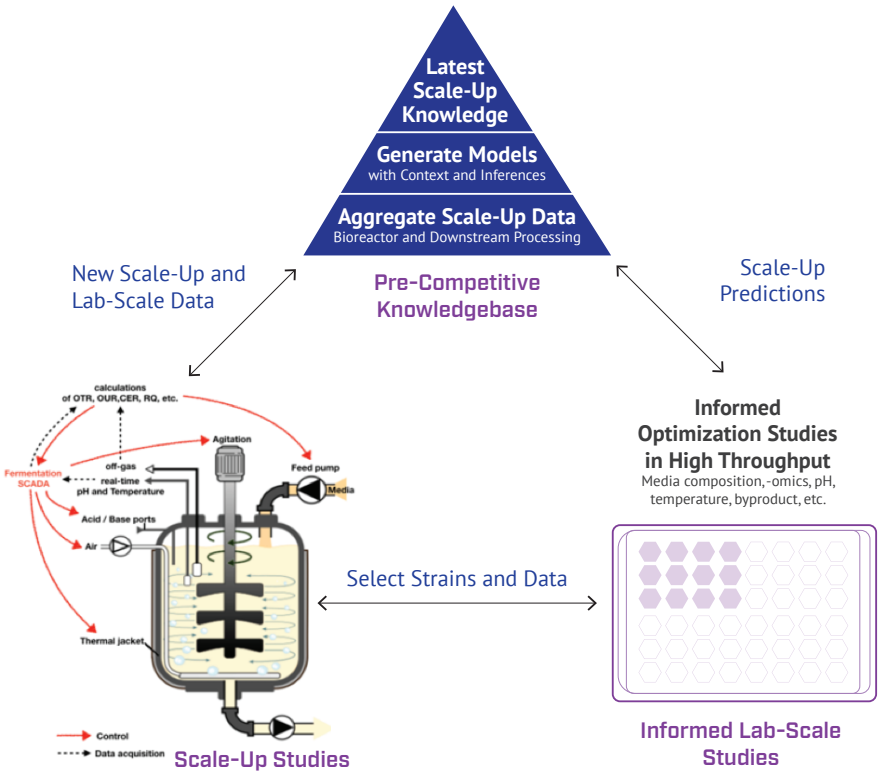

Figure 1. PCK provides a bridge between scales to share data and models and is updated with scale-up data often and therefore can be reliable to inform new lab-scale studies.

practitioners from the basic and applied realms. The PCK can serve as a key asset in this process by providing access to data that can drive decision-making and reduce (and potentially prevent) work that has already been performed (and in some cases been unsuccessful) or non-scalable solutions.

In designing the PCK, its utility can be of maximum value when it can be used to bridge the gap between foundational basic research and application towards commercialization (**Fig. 1**). Therefore, the structure of the PCK needs to reflect this. It is essential that standard terminology be used for effective communication and collaboration so that ambiguity is avoided. A PCK lexicon for biomanufacturing that includes those elements shown above should be developed and deployed and adhered to by those using the PCK across all aspects of usage of the PCK. Furthermore, having clearly defined processes for agreeing on the lexicon, for data submission, access control, curation etc will be necessary to streamline operations and usability. A PCK Steering Team should be established to provide such oversight and make decisions for the PCK (**Fig. 7**). Ultimately, this effort to standardize nomenclature and data sharing methods will help have benefits far beyond the PCK itself by standardizing the way we communicate, the way we collect, record and document data and the way that we work together.

The PCK can be used to foster community building through facilitating dialogue, partnerships, data sharing and collaborative problem solving. As the PCK is developed, listening sessions, workshops, and user-centered design processes should be employed to identify the most pressing needs that the PCK can help address. The PCK will need to have mechanisms whereby users can pose questions or needs such that the PCK community

can work together to address these. This will help identify common challenges and those that are most prevalent in advancing biotechnology, and where the PCK can be of most value. An online discussion forum should be established to enable such discussions that can be grouped by themes or by the common challenges. Whilst this will need moderation, the forum will be an essential element of the PCK.

The PCK can also serve as an educational tool to integrate and disseminate knowledge across the entire biotechnology value chain. As the PCK is developed, it would be vital to include educators who can provide insights on how the PCK can be used for educational purposes and how workforce development can be built around the PCK, such as students being given scale-up problems to generate data that can be added to the PCK. It can also serve startups who may not have the knowledge of what it takes to translate their product(s) to market due to commercial and regulatory processes and requirements.

The PCK user community should be diverse including those in academia providing the foundational knowledge to build from to those in the startup realm seeking to commercialize their first products wishing to learn successes and avoid pitfalls to those that are already seasoned in successfully bringing forward products to market and who can be a ready source of knowledge and experience. There will be different use cases of the PCK for different Bioindustrial Manufacturing Readiness Levels (BioMRLs)[14] (**Fig. 2**). The platform should be tailored to accommodate the specific needs and requirements of different users operating at various BioMRLs, and most importantly, at the higher BioMRLs where there is a sense that data and knowledge is more proprietary. e.g. operating costs of a manufacturing facility, containment measures. As discussed below, the PCK can bring most value in bridging the gap between early stage R&D efforts (BioMRL1-3) and late-stage manufacturing (BioMRL8-10), the so-called "valley of death" for bioprocesses. One such specific use case is the sharing of base media recipes for common biomanufacturing hosts (i.e. *Bacillus*, *Escherichia coli*, *Saccharomyces cerevisiae*) that result in robust, high-cell density growth at scale. Currently, many early-stage companies have to invest heavily in media development to achieve a commercially relevant process, but the sharing of these non-proprietary recipes can shorten development timelines and reduce the chances of failure in the "valley of death".

Beyond PCK users who work at different phases of product development, the PCK needs to cater to the international community. Many companies have a global presence where research, manufacturing and operations are spread over multiple geographies. Regulatory requirements vary between jurisdictions and this should be realized within the PCK. There may also be regulations on how data is shared that may influence how the PCK is structured. Additionally there are complexities in terms of language and culture and the PCK needs to be intentionally designed in a way to break down these barriers i.e. having a standard terminology lexicon. There needs to be transparency in the expectations of how PCK users join, contribute and make use of the platform, as well as clearly outlining violations and consequences. A PCK code-of-conduct will need to address these aspects and be managed within the overall governance of the PCK.

One of the key areas that the PCK can enable is streamlining the understanding and meeting the requirements of regulatory processes, especially those for agriculture, and food/feed. Relevant

14 Smanski, M. J., Aristidou, A., Carruth, R, Erickson, J., Gordon, M., Kedia, S. B., Lee, K. H., Prather, D., Schiel, J. E., Schultheisz, H., Treynor, T. P., Evans, S. L., Friedman, D. C., and Tomczak, M.. 2022. Bioindustrial manufacturing readiness levels (BioMRLs) as a shared framework for measuring and communicating the maturity of bioproduct manufacturing processes. *Journal of Industrial Microbiology and Biotechnology*. Volume 49: 5 *https://doi.org/10.1093/jimb/kuac022*

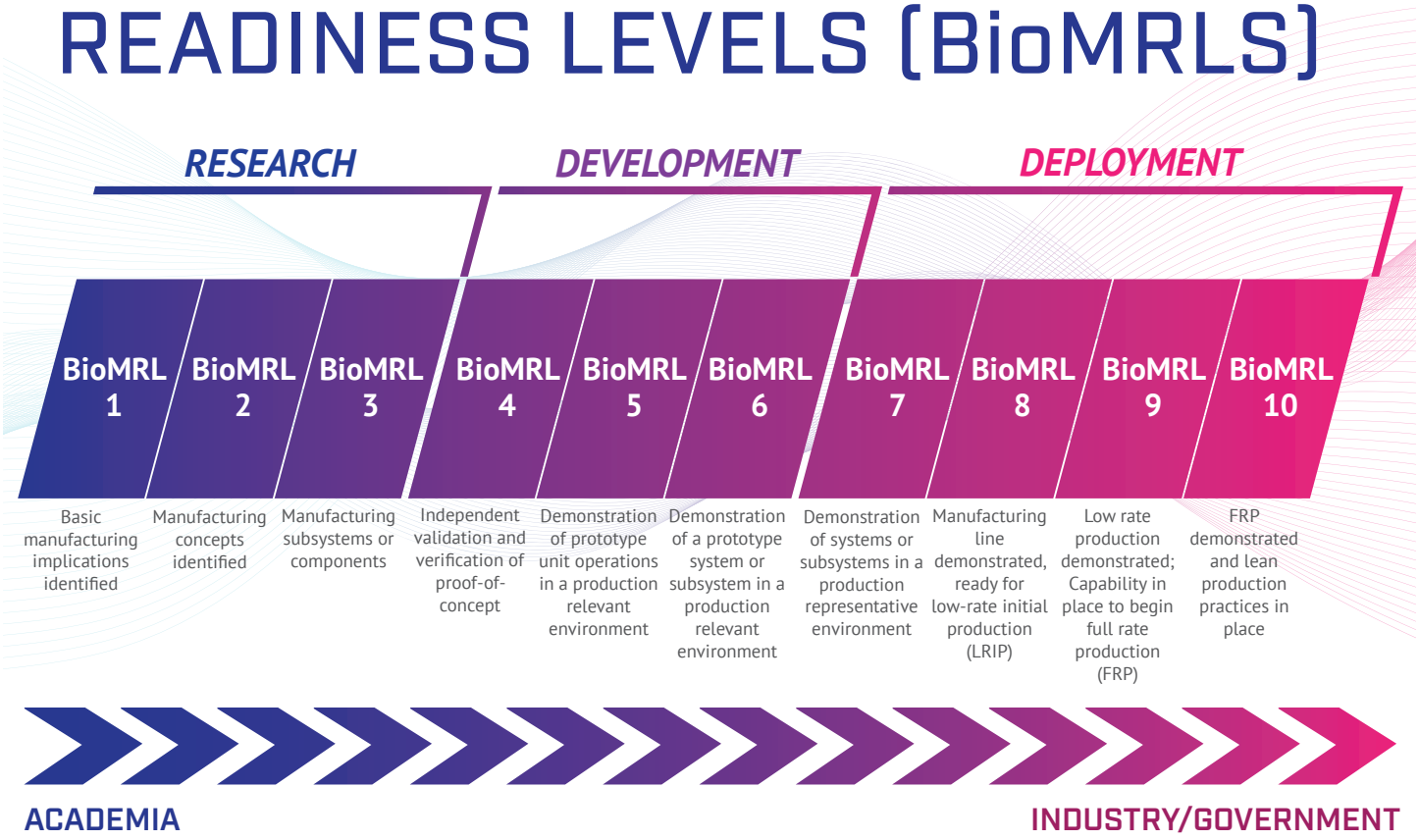

Figure 2. Biomanufacturing Readiness Levels (BioMRLs) indicating the stages for bio-based product development. The PCK can play a key role in bridging the gap between BioMRL levels 3 and 8, from Research to Deployment.

information pertaining to regulatory requirements for different types of products and different regions would be a section of information worthy to include in the PCK, and to use the embedded search functionality to make it easier for users to locate the information they need. This would require regulatory experts being willing to share their knowledge and experience. The PCK can facilitate the creation and/or deployment of standard terminology in industry that can be used in regulatory processes, and more broadly across biomanufacturing and biotechnology. This further addresses the global nature of the community the PCK would serve and aids in compliance and cross-border data sharing. It should be noted that as regulations evolve and change, the PCK will need to be updated with the current information.

In developing the PCK, there is an opportunity to learn from others what has worked and what has failed in building similar types of data systems. The key in any such system is the content and how that content is accessed and used. How users interface with PCK is critical for them to be able to get the data they need in the right format, how they can interact with other members of the user community, and with the value they derive from PCK catalyzing additional users and data providers. Through user listening sessions and workshops, these critical characteristics can be articulated such that user needs are captured and included in the development process.

Theme 2: Building Public-Private Partnerships

The PCK should serve as a bridge between foundational and translational research and between academic researchers and scientists at companies. Industry should partner with academics and government to provide the practical specifications of how to apply basic research in the commercial setting and the format, type and quality of data that will propel commercial success. The PCK also needs to be flexible enough to accommodate new technologies, types of data and models of knowledge sharing (e.g. artificial intelligence). The PCK can also be seen as a broader problem solver; once it has been used to solve one challenge, those solutions may solve others.

It is recognized that academics are not driven or incentivized to partner with industry, and therefore there needs to be mechanisms whereby they are rewarded to partner. Few academics have a steady stream of revenue, so funding for partnering and data sharing is an obvious mechanism. While academia is generally inclined to share data, the process of data sharing needs to be simplified to encourage researchers to contribute to the collective knowledge pool more readily. Barriers to knowledge sharing are being lowered e.g. pre-publication sharing through Biorxiv[15] and sites such as Halo that enable collaboration[16]. The sharing of data from academia could stimulate additional academia-government-industry partnerships since it will be easier for industry to identify appropriate labs for future work. This in turn may lead to additional funding streams for those academic groups that wouldn't otherwise be accessible without industry engagement. Programs such as NSF GOALI are already established: "GOALI is a type of proposal that seeks to stimulate collaboration between Institutions of Higher Education (IHEs) and industry. Under this proposal type, academic scientists and engineers request funding either in conjunction with a regular proposal submitted to a standing National Science Foundation (NSF) program or as a supplemental funding request to an existing NSF-funded award"[17] and other Federal funding agencies have similar programs. The PCK could serve as another mode of engagement or serve as an intermediary.

In addition to partnership with industry, student engagement is another mode whereby PCK can advance biomanufacturing and biotechnology. Academia has the key role in preparing a skilled workforce that is well-versed in data sharing practices and principles. Students could be required to engage in data curation as part of their training. FAIR data principles could be taught more broadly as part of a curriculum on data sharing and management and the PCK could serve as a data repository for example data sets and a venue for students to publish their work.

Industry also needs to be incentivized to participate in the PCK and incentivization mechanisms are discussed below (Theme 5). Industry may seek to get information faster and in a format that they can readily use. Startups may also want to learn paths to success (or failure) faster and would be incentivized through having access to information at a time when they need to make key decisions about direction, partnerships or funding. They can also use the PCK to gauge the novelty of their process against those that have already been attempted.

Federal funders can play a pivotal role in promoting data dissemination and sharing and already there are multiple government-funded data repositories.[18] However, to ensure the effectiveness of this effort, it is crucial for them to

15 *https://www.biorxiv.org*
16 *https://www.halo.science/scientists*
17 *https://www.nsf.gov/eng/eec/goali.jsp*
18 Vision, Needs, and Proposed Actions for Data for the Bioeconomy Initiative. 2023. NSTC. *https://www.whitehouse.gov/wp-content/uploads/2023/12/FINAL-Data-for-the-Bioeconomy-Initiative-Report.pdf*

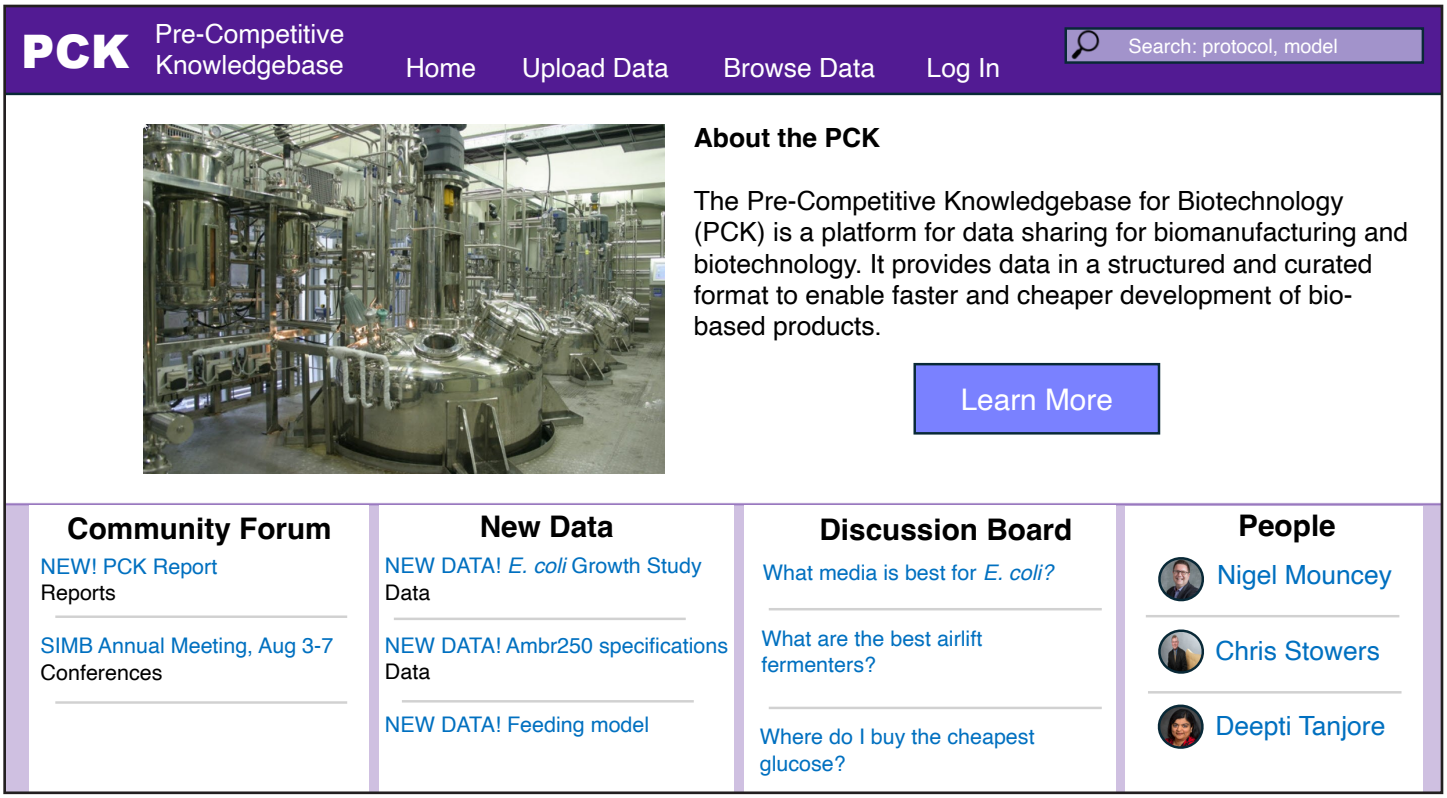

Figure 3. Mock PCK landing page that provides access to core functionality, discussion board and key contacts.

invest in the development of standards, training and incentivizing the appropriate use of metadata. This step is essential to maximize the utility of shared data through linking datasets within and across protocols. Funders can also fund tools and resources that can prepare data for ingestion into the PCK, which is essential to ensure that the PCK captures as much data as possible. As with any government funded activities, available resources are dependent on funding levels. There are also challenges related to the lack of long-term funding that is crucial for the sustained success of such programs. Computational facilities are seen as infrastructure and maintained and upgraded but data is not and therefore the longevity of data is vulnerable. Therefore, if the PCK is realized, a plan for long-term funding must be developed from more sustainable sources e.g. philanthropy.

The PCK can also serve as a broker for building communities. The PCK site should include links to relevant news stories, discussion forums and the means to connect with other users (**Fig. 3**).

Discussion forums can be set up on topics to focus on and bring those working on similar problems together to find solutions or prioritize data in the PCK that will build community. Users could post needs and ask questions. As part of the development of this report, SIMB worked with Herzog & Schindler to pilot a Polyplexus[19] discussion board on key questions arising from the workshops. These discussion boards can provide a means to collect feedback from users on PCK performance and features they would like to see. As humans are considered to be better at intuition and imagination, the boards can also offer an avenue for users to discuss and debate AI-based analyses of PCK data, when such discussions are warranted. Secondly, and as described elsewhere in this report, emphasizing user-centered design is invaluable when creating a data-sharing platform. This approach not only ensures the usability of the platform but also fosters a sense of community among its users. Thirdly, development of case

19 *https://start.polyplexus.com*

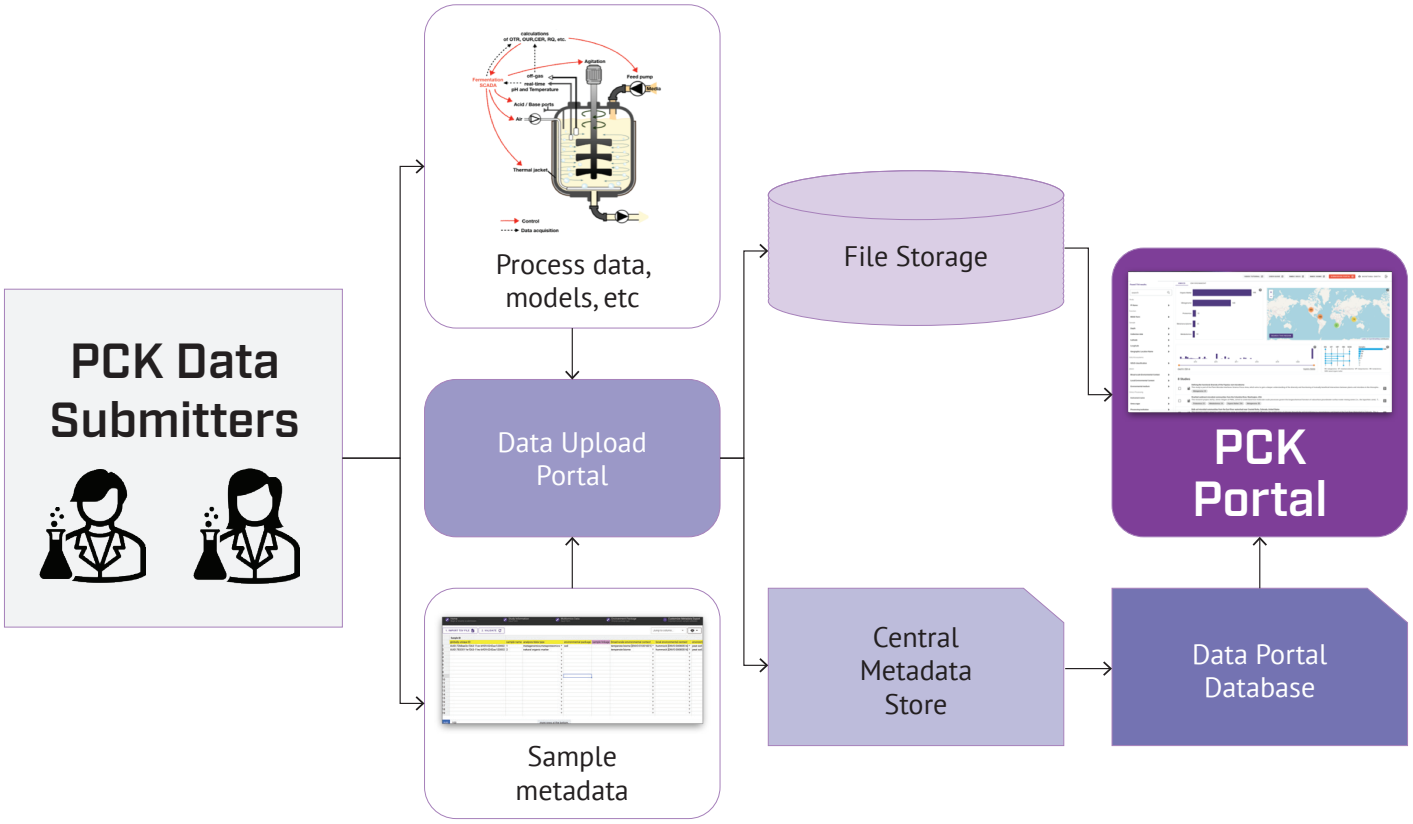

Figure 4. Proposed structure and data flows of the PCK. Credit: The National Microbiome Data Collaborative.

studies which can serve as practical examples and best practices for future projects can entice people to use the PCK.

However, challenges exist in building public-private partnerships. Universities almost all have commercialization/technology transfer offices that can overvalue data and technology despite their best intentions in moving it towards the applied arena. Integrating data across different data types and disciplines is not trivial and is made more complex in integrating data from industry and academic sources, from different types of funding.

Theme 3: Structuring the PCK

The PCK would serve as both a data repository and a data facilitator platform. As a data repository, the PCK would house data types that are of most value to those seeking to accelerate development of bio-based products. This data needs to be structured in a manner that allows for its search, retrieval, curation, updating and integration with other data types (**Fig. 4**). The data should have appropriate metadata and provenance information associated with it, and where possible, data standards employed (either existing or new) to allow for interoperability and reuse of the data, in accordance with FAIR data principles. As a data facilitator platform, it is vital that the data be accessible to all users and searchable ideally using natural language to query if the data were to be standardized and structured in such an enabling manner. The PCK should cater to a broad set of users and so the interface and tools should be easy and straightforward to access and use.

Biomanufacturing and biotechnology are broad areas in their scope ranging from chemistry, biology, engineering as technical disciplines to logistics, regulations and policy, packaging, sales

Data type	Submitter	User	Example use cases	Priority
Organism cultivation conditions	Academia, Industry	Academia, Industry	Optimizing growth conditions during scale-up	Medium
Omics data (genomics, transcriptomics, proteomics)	Academia, Industry	Academia, Industry	Identification of genes for modification for desired phenotypes	Medium
Metabolite information	Academia, Industry	Academia, Industry	Identification and quantification of known and unknown metabolites	Medium
Bioreactor specifications and operational modes	Academia, Industry	Academia, Industry	Optimization of reactor design to meet desired properties	High
Fermentation operations, feeding regimens, control points	Academia, Industry	Academia, Industry	Optimization of parameters to meet production targets	High
Downstream processing operations, recovery methods	Academia, Industry	Academia, Industry	Optimization of processes to maximize yield	High
Raw material specifications, pricing	Industry	Industry	Media optimization at scale	High
Other equipment specifications, settings, operations	Academia, Industry	Academia, Industry	Optimization of equipment settings to meet desired needs	Low
Safety, containment	Academia, Industry, Government	Industry	Safe operations to meet regulatory requirements	Low
Regulatory information	Academia, Industry, Government	Industry	Optimization of product and process specifications to meet regulatory requirements	Low

Figure 5. Potential PCK data types, data submitters and users, and example use cases. Priority is for incorporation of the data type into the PCK.

and marketing, and consumer purchasing and use. To cover all of these areas in the initial Minimal Viable Product (MVP) would be a huge undertaking that also is subject to broad situational facets that are hard to build general models from. Therefore, it is preferred to initially prioritize the content of the PCK to the technical data types associated with scale-up of fermentation and bioprocessing, such as engineering, fermentation media, operational conditions, product recovery methods and microbial physiology (**Fig. 5**). This data would take the form of processed data rather than raw data but details of data processing approaches should be shared to enable interoperability. The PCK could house developed models that can be deployed on new PCK user datasets, such as kinetic fermentation models or mixing models. Even so, there are broad data types to consider that will cater to different needs and these would be the subject of further development activities. The PCK can play a unique role in facilitating data provision that is combined from multiple sources. For example, bringing together data from historical and current projects can illustrate how biotechnological processes and capabilities evolved which may provide insights into further

optimization and successful scale-up. Tracking of decision making when developing processes is not yet widely implemented, but essential to enable any further AI that can lead to fully automated equipment, such as self-driving bioreactors and downstream processing units.

One type of data that may be easier to include is historical data. Historical data can take the form of published data from peer-reviewed articles, patents, reports, etc. Given that this data is already openly available, the PCK can serve as an aggregator and facilitator of this data bringing functional features to it such as search and indexing, to make it easier to find relevant data and sort data. As with any of the data types in the PCK, a tracking system should be established to determine how often this data is accessed.

In focusing on the core technical aspects, and to make this data FAIR, standards need to be established to ensure structure of the data, quality of the data, consistency and comprehensibility of the data. This includes standards for data quality, lexicon, defined metadata types, submission formats, metrology, physical and analytical measurement standards and fully defined equations and statistics for all calculated values. Some standards already exist e.g. those for genomics from the Genomics Standards Consortium[20]. A Bioeconomy Lexicon has been developed by the National Institute of Standards and Technology[21]. Other standards for biomanufacturing and biotechnology are in development. Where appropriate the PCK will leverage existing standards. But still others will need to be developed and it is recommended here that a standards working group is convened to work on developing these.

In principle, the PCK serves two main functions: it ingests data from uploaders and it provides data to those who need it. Knowing that all of the data initially uploaded might not be fully curated, the structure of the PCK needs to be set such that uncurated data is separable from curated data. This could be done "physically" through file structures or the web interface or it could be done through a data tagging system. Once the data is curated, it can be validated and reclassified as such. The system should also be capable of allowing users to flag data that is suspect so that it can be investigated by the curation team.

The PCK should be built with user operability in mind. As the PCK is being developed, user-centered design methods should be employed such that the modes of how users will access and work with the data are incorporated into the design of the PCK structure. Once built, further user research should be conducted to see how users are using the data and what features they would like to add and/or eliminate.

In developing the PCK, measures will need to be implemented to promote data sharing while maintaining confidentiality and protecting intellectual property. This can take the form of being able to easily search for data using natural queries (based on large language models), having a tracker that alerts when new data has been added, or when new users have joined. Verification processes must be in place to confirm that submitters have the rights to the data they share. Liability considerations should be addressed perhaps through showing a disclaimer on login that data is used at own risk. The PCK should have a flexible access model that aligns with the overarching goals of the initiative, ensuring that it serves the interests of all stakeholders effectively.

Theme 4: Data Sharing and Reporting

The PCK would be a data sharing platform and should incorporate the key characteristics for managing and sharing data to be an effective

20 *https://www.gensc.org*
21 *https://www.nist.gov/bioscience/nist-bioeconomy-lexicon*

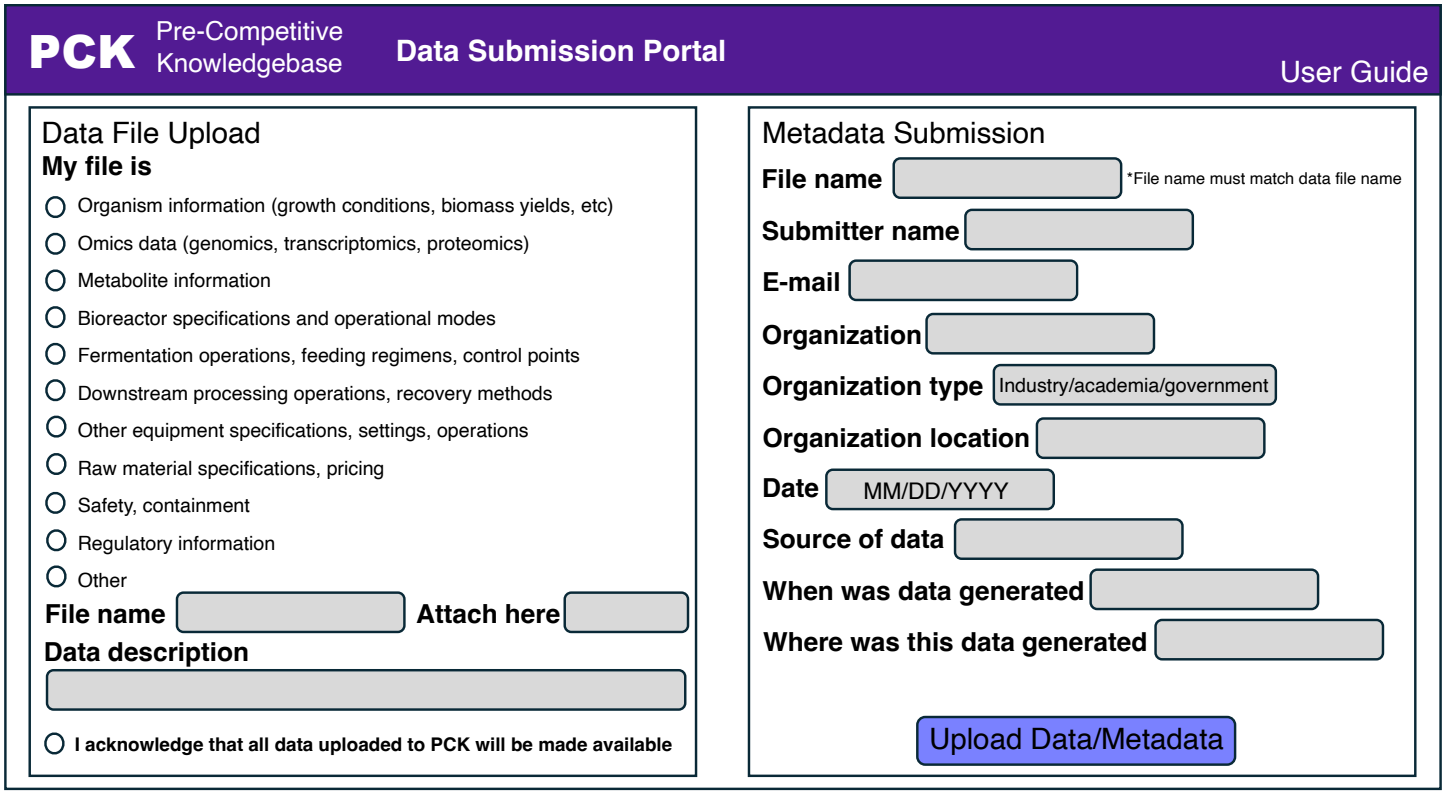

Figure 6. Mockup of potential PCK Data Submission Portal

platform that serves the needs of its user community. Using the Desirable Characteristics of Data Repositories for Federally Funded Research[22] as a guide, the PCK should be structured as follows.

Organizational Infrastructure

The PCK needs to provide straightforward access to the data so that it is readily usable to those using and contributing to it. It will be important upfront to define the roles of data users and associated permissions. There may be different tiers of users - data providers, data users and those that aid in curating the data - and this may be dependent on the model adopted for the PCK, fully open access or through a membership-type structure. There may also be tiers of users: those who pay for access that can access all information; and those that can access only a subset of information freely. If PCK adopts a membership structure, then a membership agreement will be required. The PCK needs to accommodate both start-ups and large companies, as well as the academic and not-for-profit entities. A balance should be struck between the benefits of data sharing across these and policies and practices should be inclusive of all.

The facilitation of data sharing via the PCK needs clear guidance that outlines terms of use and access. Upfront, there should be a clear understanding of who will use the data and what use cases there may be. For this reason, the PCK should keep a record of who accesses which data that can be mined in the future to identify the most valuable data sets and types. A Data Sharing Agreement may also be worthwhile to develop to give some level of protection to all parties. However, the PCK will exist mostly to facilitate open data sharing and data submitters will have to enter into the understanding that the data they

22 National Science and Technology Council Subcommittee on Open Science. 2022. Desirable Characteristics of Data Repositories for Federally Funded Research. *https://doi.org/10.5479/10088/113528*

provide may be used openly (apart from nefarious activities). Any data that is being provided from an organization (and not open sources) requires that data contributors have permission from their organization to share the data and this will be verified prior to data upload.

Digital Object Management

The usefulness of the PCK will only be as good as the quality and format of the data that is included in it. Curation will be necessary to ensure the accuracy (i.e. as much as possible, spurious data removed) and quality of the data is sufficient for use. Key to this will be structuring the data in a manner, along with metadata schema that provide standardized descriptors to enable cross-comparative and analyses of different data sets and types. The metadata schema should be able to link data of different types through e.g. standard labeling conventions. A data onboarding process should be developed that includes review, curation and approval before uploading into the PCK and data onboarding specialists to oversee this. An automated Data Submission Portal (see **Fig. 6**) could be developed that makes it straightforward for data contributors to add their data along with metadata. This contains only those fields that are most useful in the context of the PCK to make data sharing easy.

Data uploaders will be strongly encouraged to ensure all uploaded data meets the quality and formatting requirements per PCK guidelines. However, it is recognized that the requirement to reformat data can be a deterrent and limit how much data is uploaded into the PCK. For this reason, the PCK must employ substantial tools and resources needed to efficiently reformat and curate large data sets provided by users.

Where possible, the provenance of the data should be known so that credit to data providers can be given. Tracking the origin and lineage of data is helpful to ensure transparency, reliability, and accountability. Provenance information helps establish trust in shared data and it can foster future collaborations between organizations who are both interested in similar data types. However, there may be situations where it is preferable to anonymize and aggregate data which can enhance its shareability while preserving privacy and confidentiality. These techniques allow for more open sharing of information without exposing sensitive details that companies would be reluctant to share and if anonymous data uploading was permitted, companies may be more willing to share data. Regardless, datasets should be assigned digital object identifiers (DOIs) so that use of the data can be tracked for impact purposes and data can be persistently accessible if the PCK is no longer supported. As data is shared and used via the PCK, new learnings that come from cross-referencing datasets or new findings that come from their use should be added to the PCK. Tracking these learnings and their associated value will provide the business case for further development of the PCK. The value and impact of the PCK is best exemplified by use cases that have positively impacted the bioeconomy, and especially those that have contributed to the improvement of the time and cost to commercialize bio-based products. However, establishing and reporting on metrics that show e.g. company/ institution/ individual usage metrics, how much data was uploaded/downloaded in a particular period and what data types can be valuable in ascertaining the growth of the PCK and tailoring it to the needs of its user community. Depending on the funding mechanism, this may be a requirement of funding or a desire from funders.

Technology

The PCK needs a mechanism to authenticate both data submitters and data users. A simple registration system should be established so that users of the PCK can be verified (i.e. names and e-mail addresses are accurate) and that metrics

can be developed around the number of users and some limited demographic information captured.

The actual infrastructure of the PCK will be dependent on who owns and manages it and how it is developed and housed. Stability and accessibility will be hallmarks of the platform and so appropriate environments to meet both the purpose and values of the PCK should be considered. In turn this needs to have measures to prevent unauthorized access, modifications to or release of data. Included in the code-of-conduct will be the expectation that companies will not hijack the platform to conduct data transfer between them, or use the PCK as their own data repository.

Theme 5: Incentivizing Participation

Like any platform, the PCK will only be seen as valuable if there is active participation and success arising from its use. As early as possible, it will be imperative to identify early successes and develop case studies showcasing the tangible benefits and outcomes of participation in the PCK. The success of the PCK will only be realized if there is active participation by those sharing data and those using the data and for both individual contributors and companies/ organizations. Participants will need to be incentivized and incentivization may differ depending on the role of the participant. Smaller companies may be looking for information that helps them advance products to the marketplace or gain the next funding round while larger companies may be looking to access data generated by small companies that can help them in accelerating product development or market uptake. Companies may have to make business cases to their management in order for them to participate so the benefits and tangible outcomes will need to be clearly articulated. Involving companies in the development process may alleviate some of their concerns and allow them to see the benefits of the PCK early.

Incentives can take several forms such as financial incentives, a credit scheme or by demonstrable benefits. Several financial approaches can incentivize participation. Requirements for participation as part of funding awards from Federal agencies or philanthropic sources could push users into participating. The PCK could host data challenges that offer payments for data or services that can populate the PCK. Alternatively, the PCK could serve as a broker for venture capitalists to sell data packages from failed businesses. Fees for private data sharing between entities on the platform can help sustain the public aspects of data sharing, ensuring a balanced and self-sustaining ecosystem. A credit based scheme, resembling a blockchain currency, could also be implemented whereby data providers receive system credits for data deposition, and these credits can then be used to access data. For those not depositing data, such credits could be purchased to allow them data access. It should be noted though that there may be tax implications with such a model. A tiered access financial or credit model could be employed, where organizations that contribute more to the PCK should be able to access more content. A mechanism should be put in place to ensure that those who contribute the most have broader access. Data could be purchased by funding agencies or philanthropic organizations and given to the PCK to share.

Other demonstrable benefits that can help incentivize participation include establishment of new partnerships being formed between participants, any data that has helped tech transfer, attract funding, advance products, or accelerate progress. Testimonies should be collected on these successes and published. Additionally, gaining early access to technologies or products in development may be valuable. The

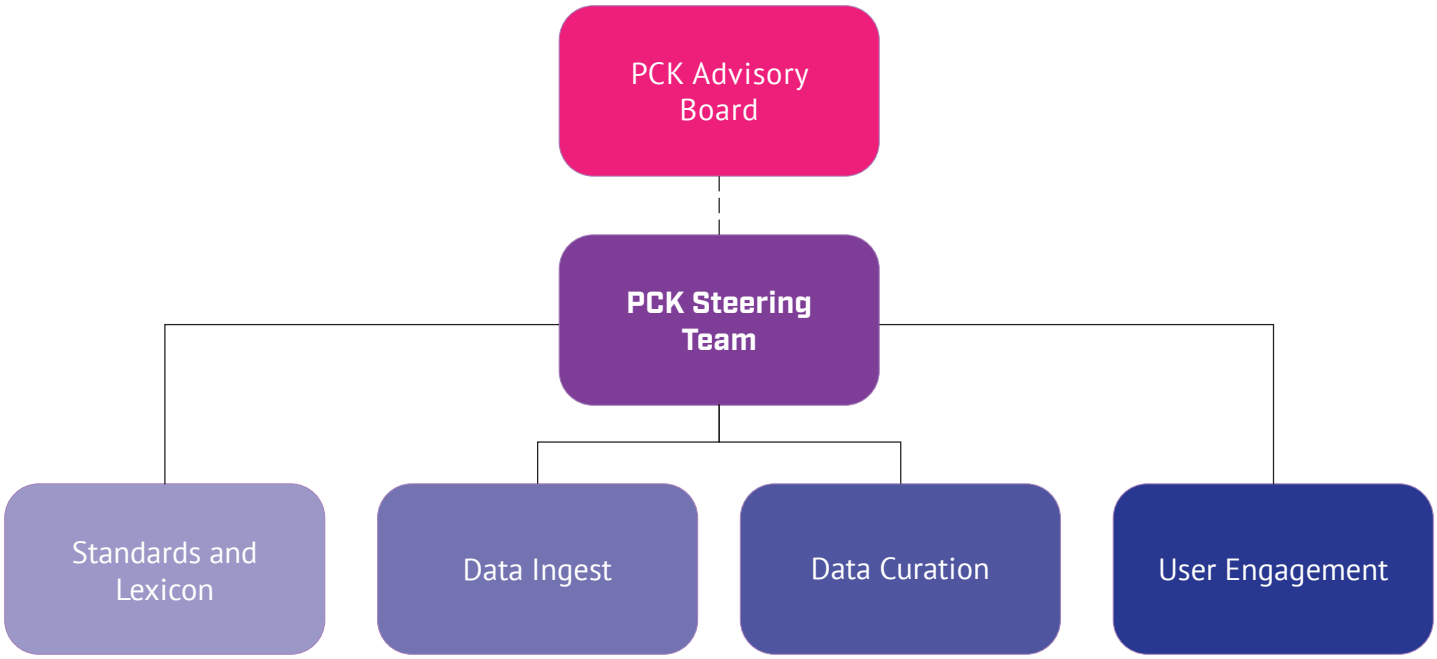

Figure 7. Proposed PCK governance and team structure. The PCK steering team would have overall oversight and be responsible for the code-of-conduct and use of the platform.

PCK could also post data needs from companies that can enable either direct company-company interaction (or with academic researchers) or identify data types that would be valuable to share in the PCK more broadly.

Another benefit that might arise from the PCK, in particular related to the standardization of data formatting and nomenclature, is helping companies achieve ISO certification that adds credibility or is a requirement for certain industries. The PCK could partner up with an ISO committee to help develop the international standards for biomanufacturing and biotechnology to which companies and their processes are assessed against. A relevant committee is ISO/TC 276: Biotechnology[23]. Through adoption of these standards within the PCK framework, companies would be able to align their processes and products to the ISO standards.

Measures should be implemented to ensure a fair and equitable distribution of costs and benefits associated with data sharing initiatives. There needs to be a mechanism that prevents users from just pulling data without contributing data to the PCK (unless the financial model reflects this). One model could be for those who do not wish or have data to contribute to serve as data curators to be able to obtain credits. Establishing a variety of approaches to contribute to the PCK can also encourage different individual personality types to contribute to the best of their ability. It is also important to take into account those institutions, e.g. Minority-Serving Institutions, who may not have the resources to contribute much data but who would benefit greatly from being able to access the data in the PCK. The data on PCK usage and other behaviors of the users participating on the platform should be protected such that strategic inquiries of companies are not revealed to the PCK host, whether public or private.

Mechanisms should be implemented that measure the impact and effectiveness of pre-competitive knowledge sharing. The numbers of users, clicks, views, interactions and usage and how these grow over time should be tracked as quantitative

23 *https://www.iso.org/committee/4514241.html*

measures. Tracking the amount of data being added into the PCK and that downloaded will also provide a measure of growth of the PCK. Testimonials, case studies, user surveys, citations and acknowledgments should all be documented. The types and diversity of organizations contributing to and using the PCK should also be documented. Interviews and discussions with users and prospective users are also valuable to conduct to determine the value they gain from the PCK, what new needs and what are the most useful types of datasets for the PCK to contain. While perhaps hard to deconvolute, the number of new companies and new products entering the bioeconomy after the launch of the PCK could be compared to that prior to the PCK. A dashboard for the impact measurements should be provided on the PCK platform.

Theme 6: Governance, Management, Ownership

The success and value of the PCK not only depends on the data and usage of that data, but also how the PCK would be owned, governed and managed. Funding for building-out the PCK can come from a number of sources e.g. Federal Agencies, philanthropy, industry or the PCK could be folded into existing efforts. In folding or linking the PCK with existing programs, there is an opportunity with the PCK to help shape those programs, ensure that complementarity and partnership happens, or leverage tools, processes and infrastructure. If funding is provided, then the PCK can be built as described above, or if there were already centralized data then the value of extending that platform could be articulated as a means to search for funding.

There needs to be a PCK steering team that is responsible for the overall design, management, maintenance and operability of the PCK, and to ensure submitters and users follow a code of conduct that outlines responsibilities, expectations and accountability (**Fig. 7**). While every effort should be made to automate the PCK processes and management as much as possible, it is expected that the steering team will need to still play a highly active role in managing the PCK to ensure data quality is maintained, operability is maximized as different data types are entered and that the participating members are abiding by the code of conduct. The overall owning body should be one that is agnostic to agenda to avoid biases and allow for the broadness of use cases. This could be not-for-profit organizations or scientific societies. If developing the PCK is managed by a third-party that may already have ongoing efforts, then management should be in accordance with that party's policies and governance structure. Additionally, clear agreements and responsibilities for data ownership and data hosting are required. These are likely dependent on where the PCK would be situated. It is certainly preferable that the PCK be managed by an entity that is able to freely disseminate information (in the case that the PCK is full open source) or can manage a subscription based model (should the PCK be operated under some kind of membership scheme).

The steering team will also be primarily responsible for ensuring the long-term funding of the PCK. Many analogous databases have been built with government funding, but ultimately erode over time due to lack of sustained funding. While the largest investment for the PCK is needed upfront, significant funding will be needed in the future to enhance the system capabilities as the data types and user set is broadened. One model is that long-term funding could be a combination of membership/ user fees as well as government funding.

As data is added to the PCK, and/or updates are made to the schema, database and/ or portal, a system is required that allows for tracking and management of data versions such that data provenance is maintained. Additionally, updates

and enhancements should be appropriately documented and archived. It is suggested that such versioning is described in accompanying documentation with the main PCK documentation. In order to track data versioning/ age, timestamps should be included to all submissions.

Data quality is another important aspect of the success and value of the PCK. Where possible, high-quality, verifiable data is preferred such that the data can be trusted. To ensure this, sources of data will need to be verified, control charting of data included, and perhaps a PCK-verified seal applied to data that meets defined quality criteria. A data standards/ quality control team should be established with the responsibility for setting specifications for the data, curating and verifying that data meets the specifications. It is expected that some of the data curation can be conducted automatically (e.g. making sure required fields have an entry) but that a final verification by a data curator be made before the data is uploaded into the PCK (**Fig. 7**). There may also be a mode whereby there are different tiers of data submissions that can accommodate varying levels of quality and relevance. This can also be based on data submission quality criteria such that some data is considered "grand-cru" in meeting all criteria, whereas other data may be of a lesser quality but still high value by meeting a fewer number of criteria. There may be additional data that meets few of the criteria but which can still be useful in bringing new insights. This data should also have guidelines and tolerance levels applied to it so that the "garbage in, garbage out" mantra is not followed. As better and more higher quality data becomes available, lower quality data should be updated or replaced. Data verification mechanisms through data curation and/or allowing users to provide feedback on posted data can help overcome this.

The PCK should follow the principles of Master Data Management (MDM) which is a comprehensive approach to managing an organization's critical data to ensure consistency, accuracy, and coherence across various systems and applications. It involves the processes, governance, policies, standards, and tools that define and manage the critical data of an enterprise, often referred to as "master data." Master data typically includes core business entities such as customers, products, employees, and suppliers but here would include the technical information described above. MDM involves creating a centralized and authoritative source of truth for master data, resolving data inconsistencies, and establishing data quality standards.

The value of the PCK will come from how it is being used, who is using it, and what comes from its use. Monitoring systems should be included within the PCK infrastructure to track which data is being accessed and how frequently, and who is accessing the data. Testimonies from those using the PCK will be critical to amplify the value of the PCK and bring others to contribute to and use the PCK. Any usage of the PCK will be subject to ensuring security of the system, the data, and usage information. One important aspect of this is to build in mechanisms that keep commercial interests confidential such that others do not obtain information around competitors' interests. This can be built into the code-of-conduct and any usage agreements that are established. Cybersecurity issues need to be addressed upfront; the PCK needs to be built in a manner that safeguards the PCK from attacks, maintains data integrity and identities of users are managed.

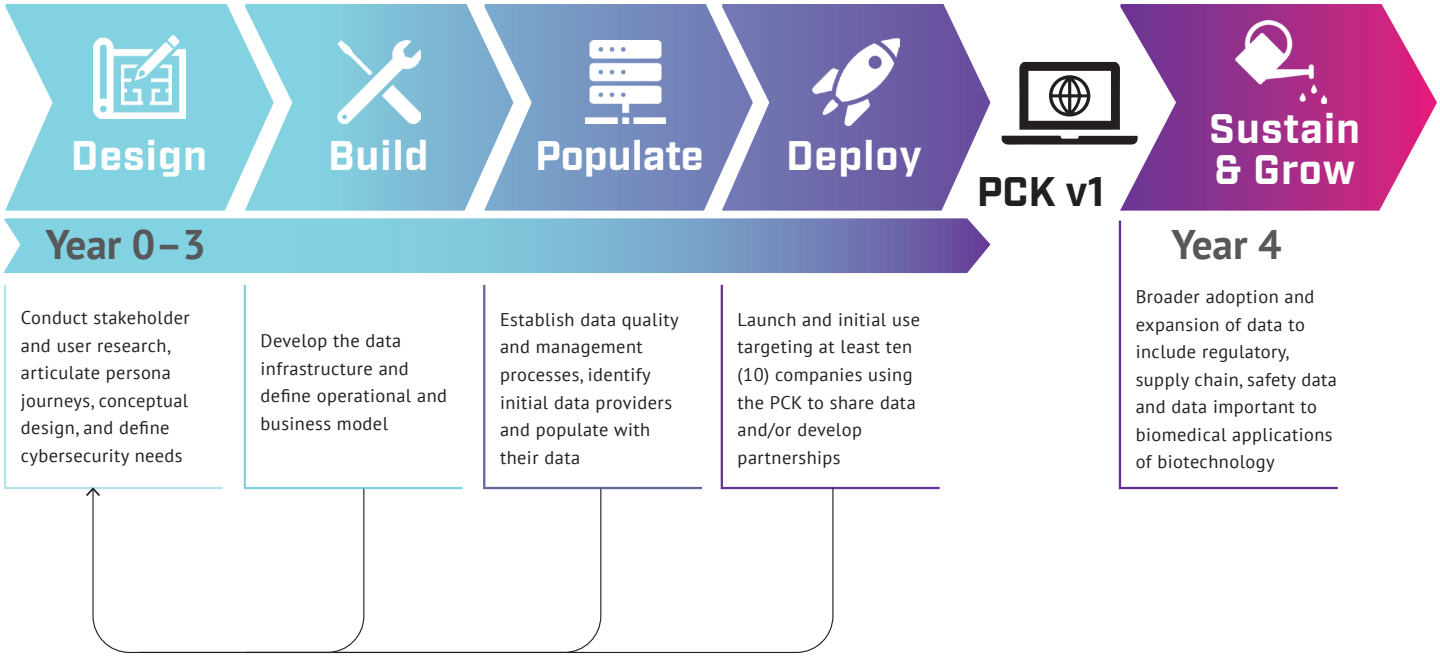

Figure 8. PCK Roadmap. A multi-phase iterative 4-year effort is proposed to generate the first version of the PCK.

ADDITIONAL CHALLENGES IDENTIFIED

For the PCK to be of value to the biotechnology and biomanufacturing community, there are additional challenges to those described above that need to be addressed:

Funding

The PCK described herein is conceptual. Funding has not yet been sought nor secured. In order for the PCK to generate value during a time period that is relevant for the bioeconomy, it needs to be developed and deployed rapidly, with an initial estimate of 4 years. Based on estimates for other data platforms, development costs could be in the range of $10M to $20M per year. Post-development phase (see below), funding also needs to be secured to grow and sustain the PCK with estimated costs of $10M/year. These estimates are based on analogous systems that require teams (20-40) people to maintain and grow the system to meet new needs.

Unintended Consequences

An important aspect of the PCK that needs to be addressed are unintended consequences: it is vital to consider the potential consequences of data sharing, including issues related to biosecurity and bioterrorism, as well as access and benefits sharing. The PCK needs also be indemnified from liabilities arising from use of PCK data. Developing safeguards and protocols to address these concerns is crucial for responsible data sharing via the PCK. It is recommended that an attorney and biosecurity experts are involved early in the development process of the PCK to address these challenges.

Buy-In and Trust

The value of the PCK lies in the data it contains and the use of that data to improve efficiencies in commercializing biotechnology products. If the data is not useful or in a format that can be used, then the PCK is not serving the community in the right manner. Therefore, continuous development of the data platform, data ingest, data structure and metadata is essential to increase buy-in, use and trust in the platform. Researchers and companies from academia, government and industry need to participate in the development process so that buy-in is achieved early in the establishment of the PCK. The PCK needs to adapt to changes in the industry to be able to address the most pressing needs of its user community.

DEVELOPMENT ROADMAP

Development of the PCK will be a complex undertaking that will require a dedicated funded team. A multi-phase approach using user-centric design is preferred so that the underlying framework and architecture is built with utility in mind from the outset. In order to deliver value and impact to the bioeconomy in a meaningful timeframe, an iterative 4-year roadmap is proposed:

- DESIGN: conduct stakeholder/user research, articulate persona journeys, conceptual design, define cybersecurity needs.
- BUILD: data infrastructure development, define operational/business model.
- POPULATE: data quality and management processes established. Data providers identified and data provision. Start with models as a lower bar data entry.
- DEPLOY: target at least 10 organizations using PCK to share data and/or develop partnerships.

At the end of this initial 4-year period, the minimal viable product will be version 1 of the PCK, a fully operational system that has already been deployed. It will be essential to show some early successes to convince funders of the need to sustain funding of the PCK and further development beyond this initial roadmap period. Strategies should be developed to support the longevity of the PCK after funding expires so that its use can persist.

Following this initial 4-year roadmap, the PCK will move into a Post-Deployment: SUSTAIN AND GROW phase that will expand the data to include regulatory, supply chain, safety data and data important to biomedical applications of biotechnology. At this time, it is hoped that broader adoption will happen and increased partnerships will happen, thereby increasing the value of the PCK to the bioeconomy.

RECOMMENDATIONS

The following recommendations are made in respect to the discussions held by the workshop participants:

- The PCK would provide a unique platform for the sharing and use of pre- and post-competitive data that has otherwise been inaccessible to those researchers and companies seeking to commercialize bio-based products and thus has an important role to play in the growing bioeconomy, and thus needs to be established as soon as is possible.
- Given the value that the PCK can have on accelerating bio-based product development commercialization, funds should be made available for scoping and stakeholder interviews to refine the vision and MVP ideas put forth here prior to funding the development work.
- The PCK should either be folded into the existing biomanufacturing and/or biotechnology data platform efforts or be funded independently. Joint partnerships with existing efforts could be formed at appropriate time(s).
- User-centered design should be the basis of PCK development so that diverse input can be sought, including that from educators, to ensure diversity of use is captured as the PCK is designed and built.
- A lexicon should be developed and deployed to provide standardized terminology across all aspects of usage of the PCK, especially given the diverse and international PCK user base.
- The PCK needs to be flexible enough to accommodate new technologies, types of data and models of knowledge sharing (e.g. artificial intelligence).
- Data housed in PCK needs to be structured for retrieval, curation, and updating and integration with other data types. The data should have appropriate standardized metadata and provenance information associated with it for interoperability and reuse.
- Incentivization mechanisms, including financial, must be established for those inputting and accessing data to make it worth their while.
- The PCK needs to provide straightforward access to the data so that it is readily usable to those using and contributing to it and terms for guidance and use communicated.
- It will be imperative to identify early successes and develop case studies showcasing the tangible benefits and outcomes of participation in the PCK.
- There needs to be a clear owning body that is responsible for the overall design, management, maintenance and operability of the PCK, and to ensure submitters and users follow a code of conduct that outlines responsibilities, expectations and accountability.

APPENDIX 1 – WORKSHOP PARTICIPANTS

Chair

Nigel Mouncey, *DOE Joint Genome Institute; Society for Industrial Microbiology and Biotechnology (SIMB)*

Co-Chairs

Deepti Tanjore, *Lawrence Berkeley National Laboratory and enScale Bio*

Chris Stowers, *dsm-firmenich*

Haley Cox, *SIMB*

Participants

Derek Abbot, *Amyris*

Albert Anis, *Valley DAO*

Stephanie Batchelor, *Schmidt Futures*

Kirsten Benjamin, *Pivot Bio*

Greg Benz, *Benz Technology International Inc.*

Kim Benz, *Proctor and Gamble (former)*

Crystal Bleecher, *Synonym*

David Blum, *Alloy Therapeutics*

Michael Clear, *Schmidt Sciences*

Tim Cooper, *Danimer*

Tim Davies, *Corteva*

Jim Dekloe, *Solano University*

Tim Dobbs, *Biosphere*

Andrew Eichenbaum, *BioMADE*

Emiley Eloe-Fadrosh, *National Microbiome Data Collaborative*

Steve Evans, *BioMADE*

Manon Herzog, *Herzog and Schindler*

Esha Khullar, *Cargill*

Vicki Liu, *LanzaTech*

Ramana Madupu, *U.S. Department of Energy*

Mark R. Marten, *University of Maryland Baltimore County*

Flo Mazzoldi, *Ginkgo Bioworks*

Joseph McAuliffe, *International Flavors and Fragrances*

Nandini Mendu, *North Carolina Biotechnology Center*

Valerie Reed, *U.S. Department of Energy*

Teri Schindler, *Herzog and Schindler*

Ryan Tappel, *LanzaTech*

Thomas Treynor, *R2DIO*

Kristina Tyner, *Culture Biosciences*

Wouter van Winden, *dsm-firmenich*

Sundeep Vani, *Biotech Startup*

Ouwei Wang, *Pow Bio*

Elisha Wood-Charlson, *DOE Systems Biology Knowledgebase*

Facilitators

Emmanuel Taylor, *Energetics*

Kirstin Janocha, *Energetics*

APPENDIX 2 - WORKSHOP AGENDAS

WORKSHOP 1: Minneapolis, MN • July 29, 2023

Time	Activity
7:00 AM	Breakfast and Badging
8:00 AM	Opening Session
8:00 AM	Intro and Welcome *Dr. Emmanuel Taylor, Senior Energy Consultant, Energetics*
8:10 AM	Initiative Overview *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
8:20 AM	Sponsor Opening Remarks *Stephanie Batchelor, Director, BioFutures*
8:30 AM	Opening Session Moderated Q&A
8:40 AM	Plenary Presentations
8:40 AM	Data Sharing Perspectives *Florencio Mazzoldi, Head of Digital Technology, Ginkgo Bioworks*
9:00 AM	Data Sharing Perspectives *Dr. Kate Sixt, Principal Director for Biotechnology, Office of the Under Secretary of Defense for Research and Engineering, Department of Defense*
9:20 AM	An Example Data Sharing System: NMDC *Dr. Emiley Eloe-Fadrosh, Lead, National Microbiome Data Collaborative*
9:40 AM	Plenary Presentations Moderated Q&A
10:10 AM	Break
10:25 AM–4:30 PM	Facilitated Sessions *Dr. Emmanuel Taylor, Senior Energy Consultant, Energetics*
10:25 AM	Introductions and Icebreakers
11:25 AM	Ground Rules for Group Discussions
11:30 AM–12:00 PM	Facilitated Discussion on Initial Sentiments and Perceptions Key Focus Questions for Consideration: ◥ What specific areas or topics could benefit from collaborative research and data sharing? ◥ What are the potential benefits of sharing precompetitive knowledge and data within the industry? (e.g. cost savings and efficiency gains from avoiding duplicative research or optimizing resource allocation) ◥ What are the potential risks or downsides to sharing PCK? ◥ What challenges are anticipated in attempting to implement the PCK?
12:00 PM	Lunch Break

WORKSHOP 1: Minneapolis, MN • July 29, 2023	
Time	**Activity**
1:00 PM	Instructions for Facilitated Discussion
1:15 PM	Facilitated Discussion on PCK Structuring Key Focus Questions for Consideration: ◥ Ideally, how would we expect users to interact with the knowledgebase? (e.g. Journal paper vs. technical report vs. AI language model) ◥ What standards or guidelines should be established to ensure the quality, integrity, and interoperability of shared data? ◥ What measures can be implemented to promote data sharing, while maintaining confidentiality and protecting intellectual property? ◥ What level of information availability is most appropriate for the knowledgebase? (e.g. Publicly searchable; free with login/account; behind paywall; tiered access with membership levels; other options / approaches?)
2:45 PM	Break
3:00 PM	Facilitated Discussion on Incentivizing Participation Key Focus Questions for Consideration: ◥ What incentivizes should be developed to encourage participation in data sharing initiatives, both for individual contributors/champions, and for companies/organizations? ◥ What measures can we implement to ensure a fair and equitable distribution of costs and benefits associated with data sharing initiatives? ◥ What mechanisms can be put in place to measure the impact and effectiveness of precompetitive knowledge and data sharing efforts? ◥ What are the challenges to participation? Are there specific incentives that can be developed to resolve each challenge identified?
4:30 PM	Recap of Findings *Kirstin Janocha, Senior Energy Analyst, Climate and Resilience Group, Energetics*
4:45 PM	Closing Remarks and Next Steps *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich* *Deepti Tanjore, Director, Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory and enScale Bio*

WORKSHOP 2: Naples, FL • Oct 27, 2023	
Time	**Activity**
7:00 AM	Breakfast and Badging
8:00 AM	Welcome and Introduction
8:05 AM	Intro and Welcome *Dr. Emmanuel Taylor, Senior Energy Consultant, Energetics*
8:15 AM	Opening Presentation *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
8:30 AM	Opening Session Moderated Q&A
8:45 AM	Plenary Presentations
8:45 AM	Plenary Talk 1
9:05 AM	Plenary Talk 2
9:25 AM	Plenary Presentations Moderated Q&A
9:45 AM	Break
10:00 AM–4:30 PM	Facilitated Sessions *Dr. Emmanuel Taylor, Senior Energy Consultant, Energetics*
10:00 AM	Introductions and Icebreakers
10:45 AM	Instructions and Ground Rules for Facilitated Discussions
11:00 AM	Facilitated Discussion on 1) Building public-private partnerships and 2) Developing mechanisms for engagement
12:30 PM	Lunch Break
1:30 PM	Facilitated Discussion on 3) Accelerating the translation of basic and applied research and 4) Streamlining tech transfer
2:45 PM	Break
3:00 PM	Facilitated Discussion on 5) Data Sharing and Reporting and 6) Managing Intellectual Property
4:30 PM	Recap of Findings *Kirstin Janocha, Senior Energy Analyst, Climate and Resilience Group, Energetics*
4:45 PM	Closing Remarks and Next Steps *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich* *Deepti Tanjore, Director, Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory and enScale Bio*

WORKSHOP 3: Columbia, MD • January 22, 2024

Time	Activity
7:00 AM	Breakfast and Badging
8:00 AM	Welcome to DSM and Safety *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich*
8:05 AM	Welcome and Introduction *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
8:15 AM	Recap of Workshops #1 and #2 *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
9:00 AM	Status of the Report *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
9:10 AM	Discussion: Introduction and Initial Sentiments *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich*
9:50 AM	Discussion: Theme 1: Accelerating the Translation of Basic and Applied Research *Deepti Tanjore, Director, Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory and enScale Bio*
10:30 AM	Break
10:45 AM	Discussion: Theme 2: Building Public-Private Partnerships *Deepti Tanjore, Director, Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory and enScale Bio*
11:25 AM	Discussion: Theme 3: Structuring the PCK *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich*
12:05 PM	Lunch
1:05 PM	Discussion: Theme 4: Data Sharing and Reporting *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich*
1:45 PM	Discussion: Theme 5: Incentivizing Participation *Deepti Tanjore, Director, Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory and enScale Bio*
2:25 PM	Break
2:45 PM	Discussion: Theme 6: Governance, Management, Ownership *Christopher Stowers, Senior Director, Biotechnology and Analysis, dsm-firmenich*
3:25 PM	Discussion: Challenges Identified *Deepti Tanjore, Director, Advanced Biofuels and Bioproducts Process Development Unit, Lawrence Berkeley National Laboratory and enScale Bio*
4:05 PM	Discussion: Roadmap, MVP and Recommendations *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
4:45 PM	Wrap-Up and Next Steps *Dr. Nigel J. Mouncey, Director, DOE Joint Genome Institute, Lawrence Berkeley National Laboratory*
5:00 PM	Adjourn

Presented by

SIMB | Society for Industrial Microbiology and Biotechnology

Visit us at *www.simbhq.org*