

Guiding Principles for Technical Infrastructure to Support Computable Biomedical Knowledge

Jamie McCusker^{1,*}, Leslie D. McIntosh^{2,3}, Chris Shaffer⁴, Peter Boisvert⁵, James Ryan⁶, Vivek Navale⁷, Umit Topaloglu⁸, and Rachel L. Richesson⁵

¹Rensselaer Polytechnic Institute, Computer Science, Troy, NY, US

²Ripeta LLC, St Louis, MO, US

³Research Data Alliance, NJ, US

⁴University of California, San Francisco, Library, San Francisco, CA, US

⁵University of Michigan, Department of Learning Health Sciences, Ann Arbor, MI, US

⁶Ryan Family Practice, Ludington, MI, US

⁷National Institutes of Health, Center for Information Technology, Bethesda, MD, US

⁸Wake Forest School of Medicine, Cancer Biology, Winston Salem, NC, US *mccusj2@rpi.edu

Abstract

Over the past four years, the authors have participated as members of the Mobilizing Computable Biomedical Knowledge (MCBK) Technical Infrastructure working group and focused on conceptualizing the infrastructure required to use computable biomedical knowledge (CBK). Here we summarize our thoughts and lay the foundation for future work in the development of CBK infrastructure, including: explaining the difference between computable knowledge and data, and contextualizing the conversation with the Learning Health Systems and the FAIR principles. Specifically, we provide three guiding principles to advance the development of CBK infrastructure:

1. Promote interoperable systems for data and knowledge to be findable, accessible, interoperable, and reusable.
2. Enable stable, trustworthy knowledge representations that are human and machine readable.
3. Computable knowledge resources should, when possible, be open. Standards supporting computable knowledge infrastructures must be open.

Keywords: Computable Biomedical Knowledge, FAIR, Open Systems

Introduction

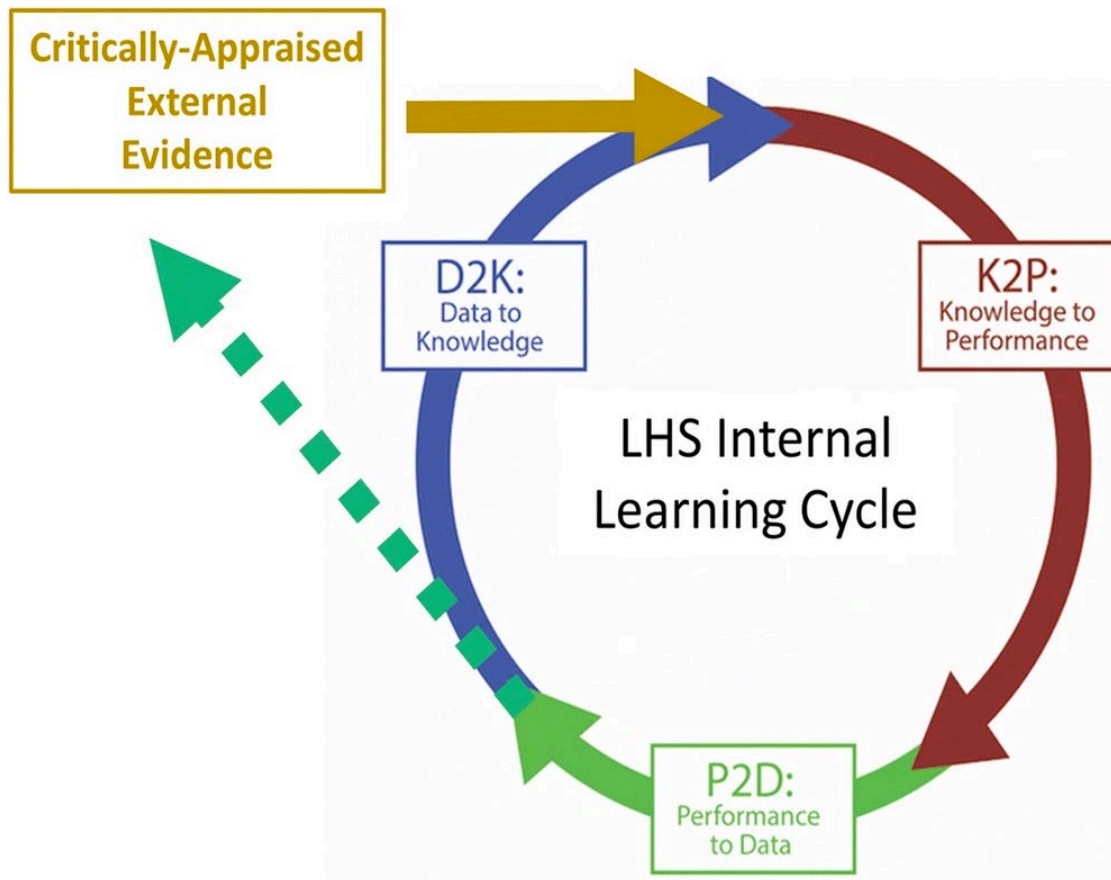
Computable formats for biomedical knowledge are needed to support the continuous incorporation of evidence and emerging knowledge structures into the Learning Health System (LHS). (Guise, Savitz, and Friedman 2018) The growth of computable biomedical knowledge (CBK) and related opportunities is driving renewed interest in knowledge

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/lrh2.10352

representation to support CBK and the infrastructure required to disseminate and apply it to different settings. However, there are many unanswered questions and opportunities for exploration.

To fully explore the requirements for a CBK infrastructure we need to understand the lifecycle of CBK, specifically how it is created, maintained, evaluated, and integrated into the broader technical landscape. This CBK technical infrastructure will need to support both management of CBK in its various forms and the movement of CBK into practice, integration with existing systems, and appropriate and effective use.

For five years, a number of interested experts from various biomedical domains have been meeting to articulate the field of CBK application, specify technical infrastructure requirements for applying CBK at scale, and develop an ecosystem for the safe and efficient application of CBK in health contexts. The purpose of this paper is to present the outcomes of these discussions and articulate the principles necessary to support this critical dimension of the learning cycle (Figure 1) infrastructure. Our group identified the principles needed for knowledge to be FAIR (findable, accessible, interoperable and reusable),(Wilkinson et al. 2016) with a focus on interoperability, and generated a number of scenarios to illustrate the importance and relevance of these principles. We extend the earlier work to the domain of computable knowledge representation to bridge the gap between the principles and knowledge implementation pathways by identifying infrastructure characteristics that can lead to the dissemination of trustworthy, open knowledge objects within an LHS. We call this approach FAIR+TO.



The learning cycle represents a dynamic coupling between data, knowledge, performance and learning community, to illustrate a continuous feedback loop that requires essential technical infrastructure support. Previously published in (Guise et al. 2018)

The MCBK Technical Standards and Infrastructure Working Group is a self-selected, volunteer interest group led by two co-chairs (originally as Technical Infrastructure under LM and CS, now JM and Bruce Bray¹) for the past five years and is supported by the University of Michigan's MCBK team. The community is open and has a listserv. The MCBK SI group has met in person three times at public meetings hosted by NLM in the summers of 2018, (Greenes et al. 2018) 2019, (Richesson et al. 2020) 2020, (Williams et al. 2020) and has had multiple phone meetings in between. In the summer 2020 meeting, the SI workgroup continued work on this article. An average of 15 participants joined these discussions. Here we present the culmination of these three years of discussions and identify current and future requirements to stimulate the development of CBK and its developing ecosystem.

¹ Dr. Bruce Bray is a Clinical Professor of Biomedical Informatics at the University of Utah, Salt Lake City, UT, USA.

CBK Technical Infrastructure - Components and Key Terms

“Computable Biomedical Knowledge (CBK) is the result of an analytic or deliberative process about or affecting human health, that is explicit, and therefore can be represented and reasoned upon using logic, formal standards, and mathematical approaches.” – MCBK Manifesto (Richesson et al. 2020)

CBK is related to, but distinctly different from, biomedical data. While CBK has a direct connection with data, managing and sharing it has its own unique needs. We see the following difference between data and knowledge. Data are the representations of information at a granular level while knowledge is data combined with processed information – such as attributes and relationships – to apply meaning to the data. (Guise, Savitz, and Friedman 2018) Here, a database is the storage of raw elements that have no meaning without linked context, such as metadata. However, metadata alone do not provide knowledge to data, as metadata still lack meaning and need interpretation. A knowledge base, therefore, will house human and machine-readable data objects, attributes, linkages, and interpretations that represent usable knowledge.

For computable knowledge to be used at scale and openly, infrastructure is needed. We use the definition of infrastructure used by the Invest in Open Infrastructure initiative. (“About Invest in Open Infrastructure” 2021) CBK has a lifecycle and can be created, modified, and updated. To support this knowledge, infrastructure must have a flexible, scalable, and maintainable pipeline for continuous input, output, growth, and use of knowledge that serves as the foundation for the Learning Health System. (Guise, Savitz, and Friedman 2018) The importance of coupling human intelligence (e.g., actual clinical data from the population) improves the accuracy of knowledge objects (Figure 1). Because knowledge evolves with new data, infrastructure should be robust and agile to accommodate the changes in scale, enhance precision in analysis, and provide services to a wide range of community needs.

The developed infrastructure should support interoperability, however, the requirements for interoperability data and knowledge are somewhat different. Interoperability in data means things like aligning format, structural metadata, and semantics, such as defined by the FAIR (Findability, Accessibility, Interoperability, and Reusable) principles. (Wilkinson et al. 2016) Data itself is passive. Only supporting infrastructure can make data discoverable, linkable, transformable, shippable, archivable, and so on. While those services also apply to computable knowledge, knowledge additionally needs to be implemented, tested, integrated with external systems, updated, versioned, re-used, and trusted. A knowledge infrastructure must therefore support what it does, not just what it is – as in the case with data object infrastructure.

CBK in a Global Pandemic

The recent global COVID-19 pandemic illustrates the need for CBK infrastructure and underscores the importance of multiple perspectives on how knowledge bases can be developed and applied. In the context of the COVID-19 pandemic, the algorithms associated

with the various epidemiological models (Iranzo *et al.* 2021) are examples of CBK. During the early stages of the pandemic, data was less available and uncertainties in model predictions were significant. As the pandemic progressed more data became available, and with the inclusion of additional model parameters the confidence interval for the mortality model predictions improved (Purkayastha *et al.* 2021). The continuous iterative LHS life cycle process shown in Figure 1 enabled the refinement of knowledge bases that resulted in the implementation of evidence-based health care policy measures for various COVID-19 impacted localities. During the pandemic, the magnitude and scope of CBKs produced were enormous, from modeling to clinical and translational research and the development of therapeutics. Such a gigantic enterprise resulted in the development and application of a wide range of platforms, services, and knowledge repositories. Fundamentally, an important aspect of the technical infrastructure is to enable knowledge to be FAIR. In the following paragraphs, we elaborate on the principles that can serve as the basis for CBK technical infrastructure.

Principles of a CBK Technical Infrastructure

We posit that CBK infrastructure should follow the FAIR principles plus two more: Trustworthiness and Openness. We call this FAIR+TO. The MCBK Standards Working Group published a good overview of the sort of metadata needed to index CBK for FAIR+T (Alper *et al.* 2022), which we see as important complementary work that fits these principles well.

Principle 1: CBK should be FAIR

We subscribe to the FAIR principles within the context of CBK, (Wilkinson *et al.* 2016) therefore CBK infrastructure should enable and promote FAIR knowledge. This is especially true of knowledge used for clinical decision support.

Findable

Findable knowledge is categorized and tagged effectively such that others can discover it and relies on knowledge being available to all. Knowledge that is hard to find never competes in the marketplace of ideas. Broadly, technical design should support scalable virtual work platforms for collection, processing, analysis, visualization, and integration of knowledge object(s). Secondly, both technical and semantic interoperability must be considered earlier during the design for supporting the CBK knowledge repositories. The findability of knowledge objects should be supported by services that include unambiguous identifier schemes (such as Internationalized Resource Identifiers, or IRIs) as well as the deployment of Application Programming Interfaces to increase the reuse of knowledge objects.

While gaps exist to make CBK findable, there is momentum to fill them. For example, there is no public “knowledge” registration system of record to catalog or index available CBK. Some funding agencies and publishers have realized the importance of the findability of the “digital assets” and making data stewardship to include long-term care of such assets. For

instance, the National Institutes of Health, has recognized the need and is seeking biomedical knowledgebase solutions that are, rightfully, separate from the existing data repositories. (National Institutes of Health 2020)

Accessible

Open and accessible knowledge systems will manifest differently than data. Open knowledge systems will help people generate, access, understand, and reuse CBK with the opportunity to bring together community and stakeholders from diverse perspectives, yet there are still many unknowns. A lack of open knowledge or the lack of attention to open knowledge systems can hinder information dissemination, hamper equitable knowledge exchange, and delay decision-making. Having open knowledge systems, however, does not offer quality assurances of the knowledge objects or knowledgebase and also can muddle intellectual property rights. As a result, accessibility, openness, and trustworthiness must be considered while building CBK infrastructure, not as an afterthought.

Interoperable

Interoperability “...is the ability of different information systems, devices and applications (systems) to access, exchange, integrate and cooperatively use data in a coordinated manner, within and across organizational, regional and national boundaries, to provide timely and seamless portability of information and optimize the health of individuals and populations globally.” (Healthcare Information and Management Systems Society 2020) The Health Information Management Systems Society (HIMSS) definition focuses specifically on data. We posit that these concepts can be used to guide the development knowledge systems as well. Interoperability brings together the power of multiple systems and ideas but depends on conformance across multiple implementations. The presence of interoperable mechanisms can drive adoption of knowledge by reducing exchange barriers. Moreover, a modular, interoperable infrastructure architecture enables integration across systems for fostering the LHS.

Interoperability of knowledge will require explicit understanding across several dimensions, including values and value sets (i.e., data), logic types, and knowledge expression types. Like data, these dimensions will require development and integration of standards of data and knowledge objects, which requires collaboration, coordination, incentives, and investments. Not having interoperable systems can impede progress through many means. Duplicating knowledge across systems, divergent knowledge, inaccurate representation of knowledge, delay in information availability, and loss of situational awareness or context can all prevent integration of computable knowledge. Yet, having interoperable systems still requires checks and cautions to reveal biases and assumptions and ensure data and knowledge are relevant and sufficient for decision-making.

Reusable

Reusability ensures that knowledge persists through time in a functional and meaningful way, and that knowledge models and ontologies evolve with the greater ecosystem. Reusability also addresses temporal concerns. Knowledge built on limited data may expand

significantly as new technologies are created and matured or as new data and information become available. It is wise to ensure that systems dependent on CBK provide the user with the current state of that knowledge as well as its provenance. Depending on the techniques employed, the outputs of the machine learning models that may generate knowledge may lack reproducibility as to how they arrived at their result. This presents a problem for consuming validated knowledge prior to putting them in practice. In order to deploy in a clinical practice, fundamental activities should focus beyond creating new models and systems, and should include studying how to best repeat the meaning behind the output of knowledge generation. Yet, formidable hurdles exist to CBK reusability. They will need to address the lack of widely accepted and consumable standards as well as governance surrounding CBK.

Principle 2: Consumers need to know how trustworthy knowledge is

Trustworthiness at its core refers to something being truthful or honest as well as acting in accordance to those principles. Thus in the CBK context, trustworthiness (Lin et al. 2020) applies to both what CBK is and what it does. A core question remains: How do I know that this CBK is correctly implemented and integrated, here and now, for this use, in this context? The issue is that the CBK itself is not entirely captured by metadata or provenance, involves both people and systems, and involves quality checks like self-testing, monitoring, and logging. Thus, “trustworthiness” is complicated and provides a vast space for exploration. A CBK infrastructure that supports trustworthiness needs to be able to address questions like these:

Knowledge:

Where did the knowledge come from? Is it what I think it is? Is it being used in the right way? Did it change and does it matter?

Scale:

Can I use this CBK everywhere? Can I use it for different things? Can I use it over time, even as it changes?

Ease of creation and evolution:

Also called $K \rightarrow K'$. Can I implement appropriately trustworthy and scalable CBK, moving my model from bench to bedside? Can I share and let others extend my CBK?

Sustainability:

Can I enforce ownership of the CBK I have created? Is my IP traceable as it evolves? Can I participate in communities that support its generation and distribution? Am I compliant with applicable law and regulation?

Publishing/Distribution/Archiving:

How do I get others to find my CBK? Can I make it part of the scientific record?

CBK is never static. It is created, distributed, regulated, modified, and applied in potentially novel scenarios. Moreover, an object or infrastructure may have a degree of trustworthiness that varies depending on its application. These bullet points highlight that “trustworthiness” is about more than the scientific validity and provenance of the core

algorithm or implementation. There are of course more stakeholders than just creators (researchers) and consumers (patients). There are owners of the IP, publishers who need to sustain a distribution network, collaborators who change and extend CBK — all of them care about “trustworthiness.”

Principle 3: CBK and its infrastructure should be open

FAIR systems serve as the necessary core of CBK technical infrastructure, while openness ensures equity of access to that knowledge. Enabling open knowledge needs to build on open and transparent foundations. (“About Invest in Open Infrastructure” 2021) Much like data, we believe in having knowledge as open as possible and only closed as necessary. With government research funding agencies like the NIH requiring open release of both scientific findings (through open access publication policies) and data (through data sharing policies), the expectation of open knowledge has become a large part of the biomedical landscape. Computable knowledge should, in turn, also follow those mandates for exactly the same reasons more traditional knowledge is required to be open.

Discussion

Through our discussions in the MCBK Technical Infrastructure Working Group (and now the Technical Standards and Infrastructure Working Group), three guiding principles emerged requiring consideration for developing a sustainable CBK infrastructure: Promote interoperable systems (1) within an accessible and open environment (2) that enable stable, trustworthy knowledge representations (3).

Knowledge is created, distributed, regulated, modified, and applied in potentially novel scenarios. When we think about what knowledge does, we can envision infrastructure that helps CBK to do things. Optimizing infrastructure could make it easier to integrate into existing health information systems - that helps CBK be used and reused in ways it was not originally designed and makes it easier to trust CBK at the point of use.

As noted in the sections above, knowledge in general - including CBK - constantly evolves, never remaining static. The interconnected relationship between data and knowledge (Data to Knowledge, D2K) suggests the principles that apply to data are extensible for developing and maintaining knowledge objects. This continuum of D2K underscores the importance of system interoperability, which is essential for enhancing the value (e.g., use and reuse) of the knowledge objects. Yet, as discussed, knowledge objects and the environment in which they evolve must accommodate the needs of knowledge objects to inform the LHS. Thus, we cannot assume the environment requirements for knowledge objects will be the same as with data objects.

Moreover, as the CBK definition implies, evidence can be subjective and not objective without a full-spectrum, robust assessment of knowledge and gaps; ‘Infrastructure’ should be value neutral, yet, the mere presence or absence of it affects value. How do we address this? And, who is going to pay for infrastructure? Does this approach make things easier or harder and for whom? Who benefits (or loses)? Does it both make the impossible possible

and the possible easy? Does it drive adoption or slow it? Raise costs or lower them? These cavities of understanding pose more questions than answers.

It is important to note our distinction between open knowledge and open knowledge infrastructure in the principles. Knowledge infrastructure must provide a level playing field for access. However, there is some value in not excluding proprietary knowledge from such an environment. However, it will have to be up to the user of that knowledge to determine if the knowledge is trustable, even if it cannot be transparently evaluated. Further, some forms of knowledge, like machine learning models, are opaque in their fundamental approach, but still have value.

A complete picture of building a CBK technical infrastructure comes with risks. If we do not make knowledge findable, we risk not having knowledge used or resources misspent through duplicative efforts. Yet, when we do make knowledge findable, the objects, attributes, and relationships may be applied incorrectly (i.e., out of context) or overfitted (e.g., creating associations where there is no causality). Ultimately, CBK technical infrastructure must support both the management of CBK in its various forms and the movement of CBK into practice, integration with existing systems, and appropriate and effective use.

Acknowledgements

This work has been made possible through the contributions of many people participating in the Mobilizing Computable Biomedical Knowledge (MCBK) working groups. This paper has been written by the named authors as well as the MCBK Technical Infrastructure working group. JM was supported by National Science Foundation (NSF), National Institutes of Health National Institute of Environmental Health Sciences (NIH/NIEHS), and IBM Research. CS was supported by NIH National Library of Medicine, Truth Initiative, and National Association of Attorneys General (NAAG). UT was supported by NIH/National Cancer Institute (NCI), NIH/National Center for Advancing Translational Sciences (NCATS), NIH/National Institute on Drug Abuse (NIDA), Patient-Centered Outcomes Research Institute (PCORI), and American Association for Cancer Research (AACR). RR was supported by NIH and PCORI.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

“About Invest in Open Infrastructure.” 2021. *Invest in Open Infrastructure*. <https://investinopen.org/about/>; Invest in Open Infrastructure Initiative.

Alper, Brian S., Allen Flynn, Bruce E. Bray, Marisa L. Conte, Christina Eldredge, Sigfried Gold, Robert A. Greenes, et al. 2022. “Categorizing Metadata to Help Mobilize Computable

Biomedical Knowledge.” *Learning Health Systems* 6 (1): e10271.
<https://doi.org/10.1002/lrh2.10271>.

Greenes, Robert, Carl Lagoze, Beltran Figueroa, and Allen Flynn. 2018. “Knowledge Infrastructure Requirements for Computable Biomedical Knowledge (CBK).” University of Michigan.

Guise, Jeanne-Marie, Lucy A Savitz, and Charles P Friedman. 2018. “Mind the Gap: Putting Evidence into Practice in the Era of Learning Health Systems.” *Journal of General Internal Medicine* 33 (12): 2237–39.

Healthcare Information and Management Systems Society. 2020. “Interoperability in Healthcare.” <https://www.himss.org/resources/interoperability-healthcare>.

Lin, Dawei, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, et al. 2020. “The TRUST Principles for Digital Repositories.” *Scientific Data* 7 (1). <https://doi.org/10.1038/s41597-020-0486-7>.

National Institutes of Health. 2020. “Biomedical Knowledgebase.” <https://grants.nih.gov/grants/guide/pa-files/PAR-20-097.html>.

Richesson, Rachel L, Bruce E Bray, Christine Dymek, Robert A Greenes, Leslie D McIntosh, Blackford Middleton, Gerald Perry, Jodyn Platt, and Christopher Shaffer. 2020. “Summary of Second Annual MCBK Public Meeting: Mobilizing Computable Biomedical Knowledge—a Movement to Accelerate Translation of Knowledge into Action.” *Learning Health Systems* 4 (2): e10222.

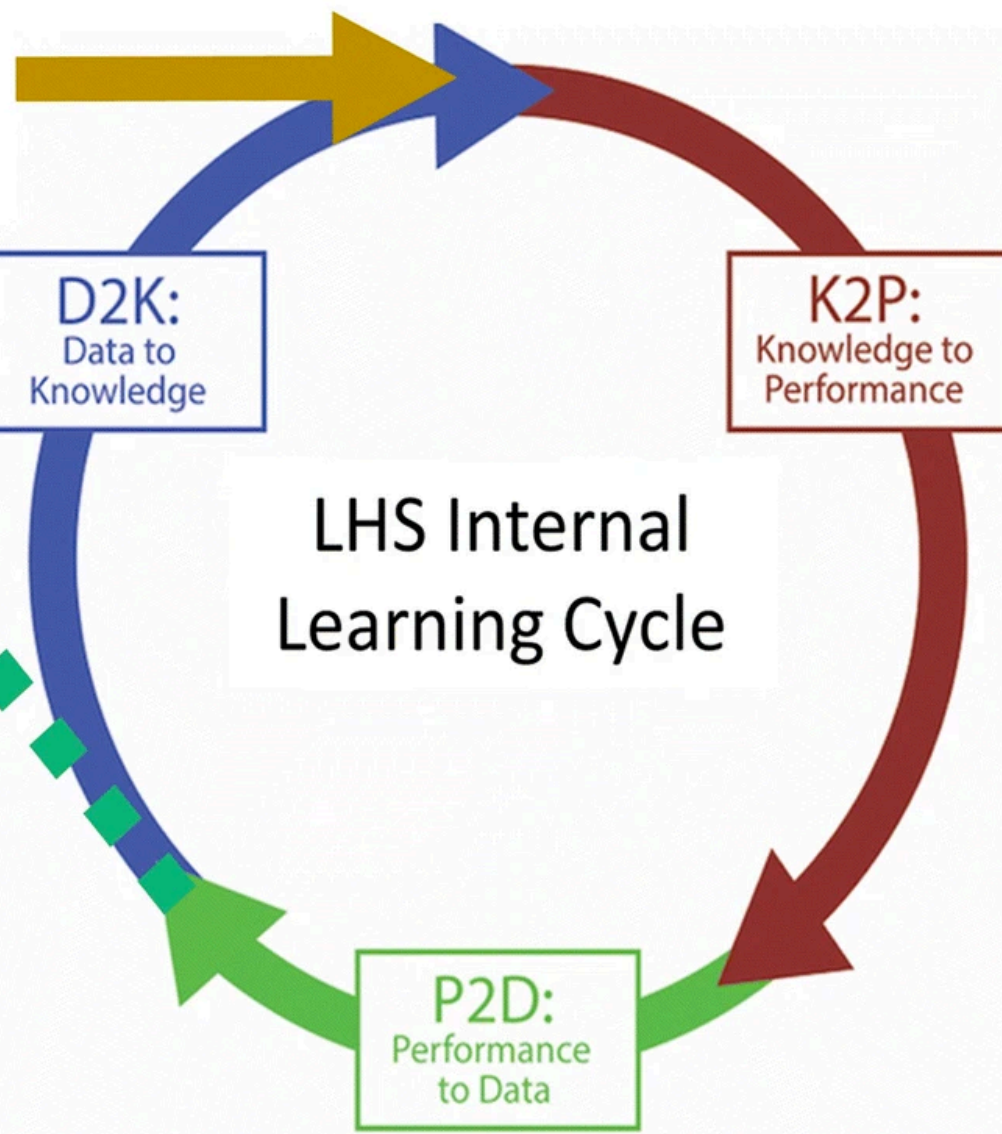
Wilkinson, Mark D, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 1–9.

Williams, Michelle, Rachel L Richesson, Bruce E Bray, Robert A Greenes, Leslie D McIntosh, Blackford Middleton, Gerald Perry, Jodyn Platt, and Christopher Shaffer. 2020. “Summary of Third Annual MCBK Public Meeting: Mobilizing Computable Biomedical Knowledge—Accelerating the Second Knowledge Revolution.” *Learning Health Systems*, e10255.

Iranzo, V., Pérez-González, S. Epidemiological models and COVID-19: a comparative view. *HPLS* 43, 104 (2021). <https://doi.org/10.1007/s40656-021-00457-9>

Purkayastha, S., Bhattacharyya, R., Bhaduri, R. et al. A comparison of five epidemiological models for transmission of SARS-CoV-2 in India. *BMC Infect Dis* 21, 533 (2021). <https://doi.org/10.1186/s12879-021-06077-9>

Critically-Appraised
External
Evidence



LRH2_10352_learning_cycle.png