Guiding Principles for Technical Infrastructure to Support Computable Biomedical Knowledge

Leslie D. McIntosh[1,2], Chris Shaffer[3], Peter Boisvert[4], James Ryan[5], Vivek Navale[6], Umit Topaloglu[7], Rachel L. Richesson[4], and Jamie McCusker[8,*]

[1]Ripeta LLC, St Louis, MO, US
[2]Research Data Alliance, NJ, US
[3]University of California, San Francisco, Library, San Francisco, CA, US
[4]University of Michigan, Department of Learning Health Sciences, Ann Arbor, MI, US
[5]Affiliation, department, city, postcode, country
[6]Affiliation, department, city, postcode, country
[7]Affiliation, department, city, postcode, country
[8]Rensselaer Polytechnic Institute, Computer Science, Troy, NY, US
[*]mccusj2@rpi.edu

ABSTRACT

Over the past three years, the authors have participated as members of the Mobilizing Computable Biomedical Knowledge working group focused on conceptualizing the infrastructure required to use computable biomedical knowledge. Here we summarize our thoughts and lay the foundation for future work in this field 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 move this field of Computable Biomedical Knowledge forward:

1. Promote interoperable systems for data and knowledge to be findable, accessible, and reusable.
2. Encapsulate knowledge objects and systems in an environment of accessibility and openness.
3. Enable stable, trustworthy knowledge representations that are human and machine readable.

Introduction

Computable formats for biomedical knowledge are needed to support the continuous integration of evidence and emerging knowledge structures into the Learning Health System (LHS)[1]. The growing appreciation of computable biomedical knowledge (CBK) is driving a new field of inquiry to better understand representations for CBK and the infrastructure required to disseminate and apply it in different settings. As with any nascent field, though, 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.

A number of interested experts from various biomedical domains have been meeting for three years to articulate the field of CBK application, specify technical infrastructure requirements towards applying CBK at scale, and to 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 learning cycle (Figure 1) infrastructure. Our group identified the principles needed for knowledge to be FAIR (findable, accessible, interoperable and reusable)[2], 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 realm of knowledge to bridge the gap between the principles and knowledge implementation pathways by identifying infrastructure characteristics that can lead to the dissemination of trustworthy knowledge objects within a LHS.

The MCBK Technical Infrastructure Group is a self-selected, volunteer interest group led by two co-chairs (originally LM and CS, now JM) for the past three years and is supported through the University of Michigan's MCBK team. The community is open and has a listserve. The MCBK TI group has met in person three times at public meetings hosted by NLM

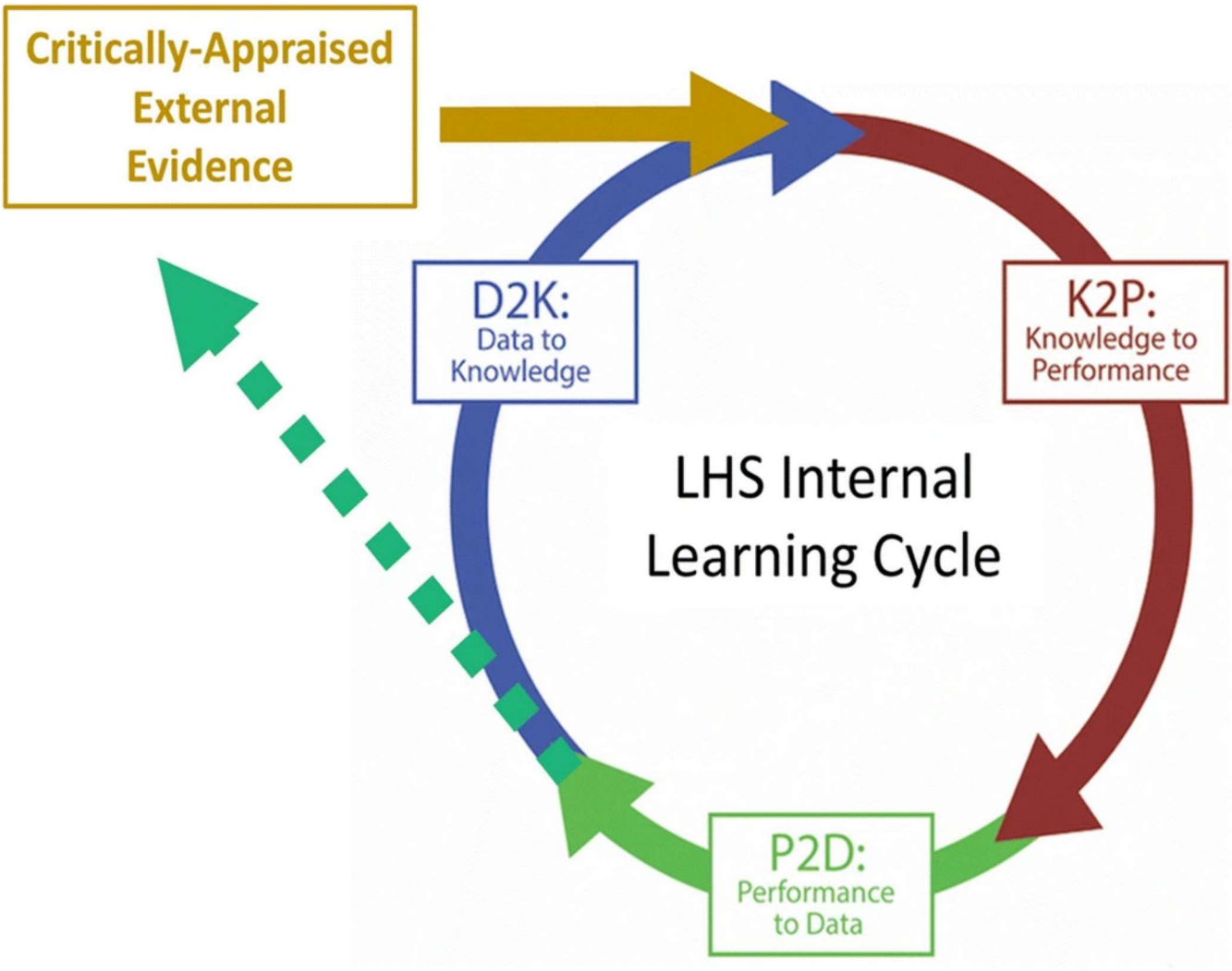

Figure 1. The learning cycle represents a dynamic coupling between data, knowledge, performance and learning community, to illustrate a continuous feedback loop that requires essential technical infrastructure support.

in the summers of 2018[3], 2019[4], 2020[5], and has had multiple phone meetings in between. In the summer 2020 meeting, the TI 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 computable biomedical knowledge and its developing ecosystem.

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[4]

Computable biomedical knowledge is related to but distinctly different from biomedical data. While CBK has a direct connection with data, we must look at CBK with prior proficiencies but support its unique needs. Accordingly, let us begin by understanding the 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[1]. Thus a database will be 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 needs interpretation. A knowledge base, therefore, will house human and machine-readable data objects, attributes, linkages, and interpretations that represent usable

knowledge.

For the knowledge in a knowledge base to be used at scale and to traverse the digital ecosystem, infrastructure is needed. As explained by the Invest in Open Infrastructure initiative:[6]

> By "infrastructure" we mean the sets of services, protocols, standards and software that the academic ecosystem needs in order to perform its functions throughout the research lifecycle — from the earliest phases of research, collaboration and experimentation through data collection and storage, data organization, data analysis and computation, authorship, submission, review and annotation, copyediting, publishing, archiving, citation, discovery and more.

Because CBK has a lifecycle and can be created, modified, and updated, a knowledgebase infrastructure must have a flexible, scalable, and maintainable pipeline for a continuous input, output, growth, and use of knowledge that serves as the foundation for the Learning Health System[1]. 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.

For optimal performance, 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 principles[2]. In some ways data is passive; thus, the infrastructure serves to 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.

The current global COVID-19 pandemic illustrates the need for CBK infrastructure and underscores the importance of a larger view and multiple perspectives for how knowledgebases can be applied. In the context of the COVID-19 pandemic, mortality predictions are based on epidemiological models. The initial model predictions assumed that social distancing will not be implemented during the early phase of the pandemic outbreak. With the governmental implementation of health hygiene and social distancing norms within a society (e.g. USA), the observed infected and mortality cases were less (one tenth) than the earlier model predictions. With community-wide data collection and sharing of actual case reports, the model was iteratively refined to revise mortality predictions. This dynamic interaction of data obtained, knowledge gained and the community interaction during a pandemic is represented by the feedback loop serving as an engine for a LHS in Figure 1.

Principles of a CBK Technical Infrastructure

The guiding principles for computable biomedical knowledge (CBK) infrastructure:

- Promote interoperable systems for data and knowledge
- Encapsulate knowledge objects and systems in an environment of accessibility and openness
- Enable stable, trustworthy knowledge representations that are human and machine readable

Principle 1: Promote interoperable systems for data and knowledge

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."[7] 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; or loss of situational awareness or context. 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.

Principle 2: Encapsulate knowledge objects and systems in an environment of accessibility and openness

Interoperable systems serve as the necessary core of CBK technical infrastructure, while accessibility and openness enhance capabilities for findability and reusability.

> "Open infrastructure is a narrower set of services, protocols, standards and software that can empower communities to collectively build the systems and infrastructures that deliver new improved collective benefits without restrictions, and for a healthy global interrelated infrastructure system." – Invest in Open Infrastructure[6]

Enabling open and accessible knowledge needs to build on open and transparent foundations. Open and accessible knowledge must be findable and reusable to enable timely, context-relevant discovery and application. Much like data, we believe in having knowledge as open as possible and closed as necessary. For even when knowledge cannot be accessible to all, how to access the knowledge and understand the knowledgebase must be transparent.

Findable

Findable knowledge is categorized and tagged effectively such that others can discover it; will depend on who and how a search is conducted; and relies on knowledge being available to all. For better or worse, 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 the Object Identifiers and Unique Identifiers schemes as well as the deployment of the Application Programming Interfaces to increase the reuse of knowledge objects.

While many gaps exist in this growing field 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 the generated knowledge. Yet, some funding agencies and publishers have realized the importance of findability of the "digital assets" and making data stewardship to include long-term care of such assets. For instance, the largest funder of biomedical research in the U.S. NIH, among the other related efforts, has recognized the need and is seeking biomedical knowledgebase solutions that are, rightfully, separate from the existing data repositories[8].

Accessible

Open and accessible knowledge objects and systems will undoubtedly manifest differently than data objects. 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 as an emerging field, there are 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. Moreover, accessibility, openness, as well as trustworthiness must be considered while building the CBK infrastructure not as an afterthought.

Reusable

Reusability ensures that knowledge persists through time in a functional and meaningful way, and that the 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 a CBK provide the user with the current state of that knowledge as well as the provenance of information. 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 in this CBK reusability process – such as the lack of widely accepted and consumable standards as well as governance surrounding CBK.

Principle 3: Enable stable, trustworthy knowledge representations that are human and machine readable

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[9] 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? That question spans more 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.

Multiple dimensions and attributes of CBK need to be consciously incorporated into the CBK infrastructure to incorporate trustworthiness into CBK infrastructure, yet we end up with more questions than answers:

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? Can I tell how it does what it does? How do we identify and incorporate the levels of trust (and biases) within the knowledge objects and pathways?

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? If the underlying knowledge evolves, is there a way to propagate that? Can I share and let others extend my CBK?

Sustainability: Sometimes thought of through the monetization or intellectual property lens. Can I enforce ownership of the CBK I have created? Is my IP traceable as it evolves? Can I easily participate in communities of practice that support generation and distribution of CBK? Am I compliant with applicable law and regulation?

Publishing/Distribution/Archiving: How do I get my CBK out there? How do others find it? How do we link it to related CBK and other resources? 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."

Narratives of CBK

Narrative 1

Contextualizing the need for CBK interoperability through a use case in a clinical setting. For instance, JR, a practicicng physician, relates:

> "When I am with a patient and we are working on a plan to control a clinical problem, we often talk about the clinical options that are recommended by clinical guidelines. This is not a new step for doctors and patients; it's as old as the written word. But at one point, the written word was a strange new thing that separated the healer from their mentor, there must have been a tension during those early generations when one trusted the text that differed from what ones teachers taught them. My generation of doctors is facing a similar transformation, rather than looking up a static set of guidelines based on a closed clinical study, we will reference computable models to inform a medical diagnosis. How I guide a patient today may not be how I would guide a similar patient next year, I can no longer trust my memory of static data. The pages of books I traded for the online text will again be traded for algorithms that exist in an ecosystem that will increase in complexity. Like the microbial world, we will be aware of its existence but we will very rarely examine these 'microbial algorithms' as we move from patient to patient.
>
> To that extent the data collected for the patient in the near term will be used to develop the algorithms that produce the knowledge objects, iteratively transformed over time for improving treatment and patient care. Not only will the computable models express knowledge-based guidelines, but a different category of computable models will need to interoperate with it to direct when these guidelines will be introduced into the clinical appointment with representation in visual form."

Narrative 2

Through the lens of CBK, finding and reusing a correct diagnosis from an biomedical informatician's perspective:

> Patient Diagnosis (DX) is one of the most important information to clinical data reuse, yet accessible structured DX data often lack accuracy in oncology. Our analysis revealed there is a statistically-significant relationship between workflow-describing variables and DX data quality. We extracted DX data from the encounter and order tables within our electronic health record (EHR) for a cohort of patients with confirmed neoplasms. We built and optimized logistic regressions to predict the odds of fully-accurate (i.e., correct neoplasm type and anatomical site), inaccurate and sub-optimal (i.e., vague) DX entry across clinical workflows. We found that differences across clinical workflows and the clinical personnel producing EHR data affect clinical data quality. The most attributed

reason was the inconsistencies in DX representation and finding the correct DX through the search within the EHR has been time consuming and many times providers give up and enter free text patient DX, which leads to many other problems. Clinicians and researchers reusing oncological data should consider such DX representation challenges when conducting secondary analyses of EHR data.

Narrative 3

An example of trusting CBK through in a clinical setting:

> When my patient asks me how sure I am that our plan is trustworthy, then I will have to give the ropes of the web a shake. The foundation of this answer will have been formed during my education starting with the basic sciences and extending through my clinical training. I will see how the data I collect and curate in my clinic is passed through the rest of the learning health system and returned to me as actionable knowledge. My understanding of the basic principles of science with a transparent path through the iterative maturing of the knowledge model will allow me to more honestly answer my patient when they ask pointed questions. This will be a novel aspect of future clinicians, they will be native to these dynamic cycles and will have a stronger intuition to the extent they are trained effectively. As far as how well a piece of fresh knowledge works when i need it to, it will probably fail when the waiting room is full and i'm an hour behind schedule. The future will still be annoying.

Discussion

Through our discussions in the MCBK Technical 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.

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 being 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, the 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.

References

1. Guise, J.-M., Savitz, L. A. & Friedman, C. P. Mind the gap: putting evidence into practice in the era of learning health systems. *J. general internal medicine* **33**, 2237–2239 (2018).
2. Wilkinson, M. D. *et al.* The fair guiding principles for scientific data management and stewardship. *Sci. data* **3**, 1–9 (2016).
3. Greenes, R., Lagoze, C., Figueroa, B. & Flynn, A. Knowledge infrastructure requirements for computable biomedical knowledge (CBK)). Tech. Rep. (2018). http://hdl.handle.net/2027.42/140738.
4. Richesson, R. L. *et al.* Summary of second annual mcbk public meeting: Mobilizing computable biomedical knowledge—a movement to accelerate translation of knowledge into action. *Learn. health systems* **4**, e10222 (2020).

5. Williams, M. *et al.* Summary of third annual mcbk public meeting: Mobilizing computable biomedical knowledge—accelerating the second knowledge revolution. *Learn. Heal. Syst.* e10255 (2020).

6. About Invest in Open Infrastructure. https://investinopen.org/about/ (2021). Accessed: 2021-02-01.

7. Healthcare Information and Management Systems Society. Interoperability in healthcare. https://www.himss.org/resources/interoperability-healthcare (2020). Accessed: 2021-02-01.

8. National Institutes of Health. Biomedical knowledgebase. https://grants.nih.gov/grants/guide/pa-files/PAR-20-097.html (2020). Accessed: 2021-02-01.

9. Lin, D. *et al.* The TRUST principles for digital repositories. *Sci. Data* **7**, DOI: 10.1038/s41597-020-0486-7 (2020).

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. The specific aspects of knowledge storage and retrieval will be addressed by the MCBK standards working group.

Author contributions statement

LM, CS, PB conceived of the framing of the paper through input of the MCBK TI working group. All authors contributed to the ideas presented. LM prepared the original draft with CS, PB, JR, VN, UT, RR, and JM adding to many versions. JM, RR, and LM completed the editing. All authors reviewed the manuscript.