Analysis of Pioneering Computable Biomedical Knowledge Repositories and their Emerging Governance Structures

Philip Sahr Amara[1†], Marisa Conte[2], Allen Flynn[1], Jodyn Platt[1], Marie Grace Trinidad[3]

[1]Department of Learning Health Sciences, University of Michigan Medical School, Michigan 48109-0624, USA

[2]Taubman Health Sciences Library University of Michigan, Michigan 48109-1382, USA

[3]School of Public Health, University of Michigan, Michigan 48109-2029, USA

Keywords: Computable Biomedical Knowledge; Governance; Repository; Trust; Sustainability

Citation: Amara, P.S., et al.: Analysis of Pioneering Computable Biomedical Knowledge Repositories and their Emerging Governance Structures. Data Intelligence 4(3), 653-670 (2022). doi: 10.1162/dint_a_00148
Received: Sept. 10, 2021; Revised: Jan. 18, 2022; Accepted: Feb. 10, 2022

ABSTRACT

A growing interest in producing and sharing computable biomedical knowledge artifacts (CBKs) is increasing the demand for repositories that validate, catalog, and provide shared access to CBKs. However, there is a lack of evidence on how best to manage and sustain CBK repositories. In this paper, we present the results of interviews with several pioneering CBK repository owners. These interviews were informed by the Trusted Repositories Audit and Certification (TRAC) framework. Insights gained from these interviews suggest that the organizations operating CBK repositories are somewhat new, that their initial approaches to repository governance are informal, and that achieving economic sustainability for their CBK repositories is a major challenge. To enable a learning health system to make better use of its data intelligence, future approaches to CBK repository management will require enhanced governance and closer adherence to best practice frameworks to meet the needs of myriad biomedical science and health communities. More effort is needed to find sustainable funding models for accessible CBK artifact collections.

[†] Corresponding author: Philip Sahr Amara (E-mail: amaphili<amaphili@umich.edu; ORCID: 0000-0002-8544-3884).

1. INTRODUCTION

Over the past few decades, digital repositories have become critically important information resources for large businesses, publishers, universities, governments, and cultural institutions. Universities especially have invested heavily in institutional repositories that promote scholarly communication and preserve intellectual works [1]. Beyond journal publications, other types of digital materials such as monographs, theses, dissertations, preprints, conference proceedings, lecture notes, and data sets are increasingly being stored in digital repositories and shared with a broader public audience [1]. Decreasing digital storage costs, the development of new metadata standards for digital artifacts, and the advent of open-source software platforms for the preservation and management of digital content have made it somewhat easier and cheaper to provide public and private digital repositories [1].

Since many scientific fields have distinctly computational branches (e.g., digital humanities or computational biology), complete scholarly communication increasingly requires sharing software along with written publications and data sets. Over the past several years, interest in curating and maintaining a wide variety of scientific software artifacts has been growing [2]. Here we focus on one aspect of this movement to share computer-processable scientific content. Our focus is on a wide spectrum of evidence-based statistical and logical models about biomedicine and human health, especially in cases when those models can either be processed or executed by computing machines, i.e., computable biomedical knowledge artifacts (CBKs). It is important to understand this emergent field to have a better understanding of how novel forms of information are made available to further enhance the information infrastructure supporting the delivery of health care.

To help explain what CBKs are and are not, we rely on Boxwala et al.'s four-level knowledge representation framework [3]. According to this framework, Level 1 knowledge representations include written papers, tables, and graphics intended for people to read. Next, Level 2 representations of knowledge include semi-structured data objects that are suitable for computer processing. Therefore, certain computer-readable configuration files and knowledge represented using various forms of markup (e.g., XML) are included in Level 2. Moving up a level, Level 3 representations of knowledge have highly formalized content providing instructions that computing machines can execute. In short, Level 3 includes procedural software code. Finally, Level 4 representations of knowledge involve procedural software code that is already deployed in one or more real-world computing environments. Level 4 knowledge representations are operable and live instances of running code [3, 4].

The repositories we examined collect artifacts exhibiting both Level 2 (computer processable) and Level 3 (machine-executable) representations of biomedical and health knowledge. Collectively, we refer to these representations as computable biomedical knowledge artifacts (CBKs). As such, CBKs include representations of statistical models, computable guidelines, executable rules, and other types of computer-processable and executable biomedical content. Our primary aim is to learn more about current approaches being taken to establish and operate CBK repositories.

2. BACKGROUND ABOUT KNOWLEDGE MANAGEMENT

The need for active, sustained knowledge management efforts in biomedicine to keep track of what the world knows about human biology and health is longstanding. Now that need is simultaneously growing and changing rapidly. Corresponding with increased global investment in artificial intelligence (AI) from $1.3B in 2010 to almost $70B in 2019 [5], biomedicine is experiencing a rapid increase in CBK related to artificial intelligence and machine learning (AI/ML) [6, 7]. To improve healthcare and achieve gains in human health, more people are producing and sharing computable analytic procedures, risk-scoring, and predictive AI/ML models, computable guidelines, clinical decision support artifacts, and other types of useful CBKs [8–11]. Meanwhile, extraordinary increases in the world's knowledge of human genetics are changing how diseases are diagnosed and treated with the help of CBKs [12]. Scientists now express hope that by using CBKs, the time lag between biomedical discovery and implementation of corresponding improvements in global clinical practices will be greatly reduced [13, 14]. Along these same lines, healthcare experts now call for more application of data intelligence to routine clinical practice [15, 16].

At present, digital repositories holding collections of shareable CBKs are novel and have not yet been widely studied [14]. Some CBKs are used in clinical practice to generate advice related to the care of individuals. These CBK artifacts are components of Clinical Decision Support Systems (CDSS). There is some relevant prior work on managing the knowledge in CDSS. It has been suggested that healthcare organizations using CDSS should form committees to review and validate CBK content and routinely analyze the relevance and value of the CBK in CDSS for advice recipients [17–19]. One study has argued for federal governance and a regulatory regime for CBK artifacts [20]. It is also clear that organizations currently implementing CDSS have developed differing governance and content management approaches [17]. The success of existing efforts to manage shareable CBK artifacts used by CDSS, such as the Agency for Health Research and Quality's (AHRQ's) CDS Connect program, ultimately depends on good governance arrangements [21].

Additionally, there is a large body of prior work on knowledge management that can help guide studies like this one [22–25]. Despite the large body of prior work on knowledge management, there is still a gap in evaluating current governance, particularly for CBK repositories, which motivates our study. To help frame our work, we limit this background discussion of advances in knowledge management to the scope of the Trusted Repositories Audit and Certification (TRAC) guidelines for digital repositories [25, 26]. These guidelines usefully summarize years of exploring and learning what it takes to manage evidence-based digital knowledge artifacts of all types over the long haul.

The TRAC guidelines are based on prior work to develop the Open Archival Information System (OAIS) Reference Model [27]. OAIS is widely recognized in the knowledge management community as the standard model to follow for better repository infrastructure [28]. OAIS' conceptual model stipulates what is generally needed to ensure the long-term preservation of knowledge artifacts and to manage an archival information system [28]. For these reasons, OAIS calls out the processes of ingesting, archiving, describing, and sharing digital knowledge artifacts. It also emphasizes the need for planning and administration to have reliable knowledge repositories. Useful though it is, OAIS offers only the highest level of guidance to repository builders [29].

Before TRAC, to extend the OAIS, the Research Libraries Group, Inc. (RLG) and the Online Computer Library System (OCLC) published the TDR, a guide to achieving Trusted Digital Repositories (TDR). The TDR lists attributes of trusted repositories, enumerate who is responsible for achieving those attributes, and establishes criteria for assessing trustworthiness [25]. Further work to refine the TDR resulted in the TRAC framework [26].

The TRAC framework outlines a set of knowledge repository-related attributes that can be assessed to measure overall digital repository trustworthiness. As shown in Figure 1 below, TRAC divides its attributes into three broad groups: Organizational Infrastructure, Digital Artifact Management, and Technology and Security. TRAC then defines, for each attribute, one or more measurable indicators for determining the degree to which each attribute has been realized by any given knowledge repository.

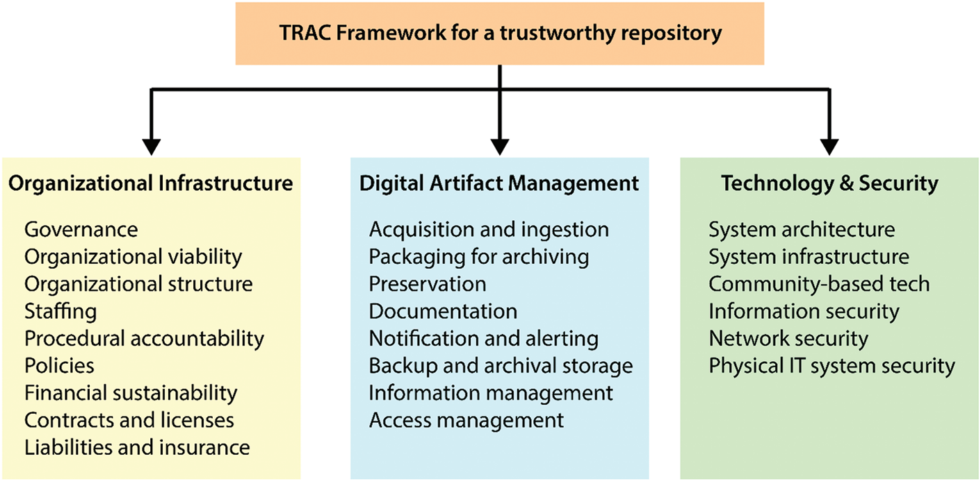

Figure 1. Trustworthy Repositories Audit and Certification (TRAC) Framework Attributes in Groups.

TRAC was not created with CBKs in mind. Instead, TRAC was formulated to assess digital repositories that preserve and share artifacts represented in human-readable formats (i.e., Level 1 artifacts on the Boxwala et al. framework outlined above). Therefore, CBK repositories may require different organizational infrastructure, technology, and security. However, standard auditing tools used to assess the trustworthiness of digital libraries, such as the TRAC, still offer valued guidance on standards, policies, and governance arrangements that may help ensure the effective and sustainable operation of any digital repository, including those holding CBKs [25, 26].

As TRAC suggests, the governance of CBK repositories is as important as their knowledge representation, ingestion, and storage processes [17]. Considering repository governance practices early in the design of CBK repositories has the potential to minimize unintended consequences and wasted effort [19].

Our aim is to describe a group of early CBK repositories and their current approaches to CBK repository governance, thereby advancing understanding of how existing CBK repositories are conceived and managed and which strategies have been employed to ensure trust, sustainability and growth. Here we examine the organization and governance of a sample of four existing CBK repositories against the relevant sections of the TRAC framework. We also examine the perceptions of operators and managers of CBK repositories about what constitutes suitable CBK repository organization and governance. Our study describes and thus reveals information about 1) the landscape within which new CBK repositories are being developed and 2) the current governance structures and strategies for ensuring trust, sustainability, and growth exhibited by this sample of CBK repositories. Since the application of data intelligence in practice involves generating and deploying certain types of CBKs, we believe CBK repositories are key components of data intelligence infrastructures [30].

3. MATERIALS AND METHODS

Our study population was a select group of existing CBK repositories in North America and Europe. Each CBK repository currently collects, manages, stores, and shares CBKs and accompanying metadata about CBKs.

We selected CBK repositories by first identifying 42 candidate repositories in consultation with experts in computable knowledge for health and health research, including the research team, as well as members and leadership of the *Mobilizing Computable Biomedical Knowledge* network (see www.mobilizecbk.org). Next, we applied the following criteria to finalize our list of included CBK repositories. To be included, each repository had to be accessible online, make some or all its CBK content publicly available, and have sufficient CBK content under management to suggest that the repository founders had previously considered questions of CBK repository governance and/or organization. Since the focus of our study is specifically on repositories that hold and manage CBKs, repositories of human-readable documents and PDFs, repositories of datasets, and more generally, digital libraries with only human-readable and non-machine processable biomedical knowledge (i.e., Level 1 or Level 2) content were excluded.

Based on the inclusion criteria, we initially identified six candidate CBK repositories. Participating CBK repositories were then recruited into our study by emails sent to listed accounts for CBK repository organizers or managers. Representatives of four of our six candidate CBK repositories agreed to participate (repositories A, B, C, D). For the four participating CBK repositories, interviews with CBK repository organizers or managers lasted one hour and were conducted by video conference between August and September 2020. All four interviews were recorded and transcribed, providing our data set.

To collect these data, we created an interview protocol with open-ended questions (supplementary material). Our questions stemmed from items in the TRAC Framework and covered organizational infrastructure, digital artifact management, technical infrastructure, and security. Before collecting data about CBK repositories, we validated our interview protocol through test interviews with managers of two digital repositories currently in operation at the University of Michigan.

To triangulate our findings, we consulted the websites for each of the participating CBK repositories to identify mission statements, organizational principles, governance documents, and other documentation relevant to the dimensions of the TRAC framework. We also searched the peer-reviewed literature for documentation of the CBK repositories and their artifacts. To uphold the anonymity of the participating repositories, we intentionally do not cite any papers that mention the four repositories studied.

3.1 Data Analysis

The recordings and transcripts from our four interviews were analyzed using an inductive approach to thematic analysis [31, 32]. We applied the TRAC framework to develop codes that were then organized and iteratively defined based on the data. The transcripts were coded by two coders using MAXQDA 2020 (VERBI Software, 2020). Relevant texts from the transcripts were independently categorized by each coder into the various TRAC codes to form coded text segments. The coders met to resolve any differences in coding. The final coded segments within each TRAC code were then analyzed to further summarize the data. Two of the three highest-level categories from TRAC (Figure 1) guided our analysis: organizational infrastructure and digital artifact management. We did not focus specifically on issues of technology and security.

TRAC's higher-level category of organizational infrastructure has sub-categories for governance, organizational structure, repository mandate or purpose, scope, roles, and responsibilities, funding, and finances, contracts, and licenses. Within the higher-level Digital Artifact Management category from the TRAC framework, related sub-categories include ingestion (acquisition and creation of an archivable package), preservation (documentation and storage), information management, and access management.

4. RESULTS

The resulting repository sample included both private and government-initiated repositories intended to address research challenges faced by certain research communities. Overall, CBK repository governance ranged from highly organized and established multi-institution partnerships to more nascent governance structures led by a small team of researchers. These smaller repositories, however, could respond more quickly to community-defined needs and were in the process of constant evolution. Well-established, multi-institution repositories could better describe their overall organization and plans due to more stable sources of funding and more standardized processes. Newer repositories, however, were more agile in their governance approach, leaving open the possibility of innovative governance and newer approaches to repository management.

4.1 Results about Organizational Infrastructure

All four computable biomedical knowledge (CBK) repositories studied were born out of a need to address gaps in existing resources or to provide structure and routine to ease the workload of researchers working in well-defined, often project-based communities (Table 1).

Table 1. Summary of results about three aspects of Organizational Infrastructure.

	Repository A	Repository B	Repository C	Repository D
The gap being addressed by the CBK repository	Lack of harmony across datasets and algorithms used for machine learning in a wide variety of domain areas	Lack of repository infrastructure to support research collaborations for many scientific projects	Lack of repository infrastructure to support machine learning and data science in biomedicine	Lack of repository infrastructure for sharing previously trained predictive models
Size of CBK artifact collection	Small	Large	Large	Small
Funding	Research grants	Research grants, institutional and foundation support	Research grants and institutional support	Research grants and institutional support

We also analyzed each repository's mission statement but do not quote those statements here to maintain repository anonymity. According to the TRAC framework, an adequate mission statement should reflect "a commitment to the long-term retention of, management of, and access to digital information" [26]. Motivations for creating the repositories expressed in mission statements included all the following: enabling reproducibility of scientific studies, publishing CBK, sharing CBK within research teams, sharing CBK across research teams, making CBK findable, accessible, interoperable, and reusable (FAIR), storing and archiving CBK, and accelerating biomedical research overall.

While three CBK repository mission statements did not describe a long-term commitment, two did describe plans to ensure future sustainability. One repository owner described their CBK repository in an ephemeral manner, anticipating better technology in the future that will render their own CBK repository no longer necessary or relevant, and noted that this determination shaped the development of the repository's infrastructure, stating: "My hunch is that much better models will be developed in the future. So, the time to live is limited [for] pragmatic reasons. […] if this were the goal, then I think the infrastructure [of the repository] would be slightly different."

Organizational Structure and Staffing: For the youngest CBK repositories, we found the organizational structure to be largely ad hoc and continuously developed in response to emerging needs. Two of the four repositories included in this study retain a small core team of developers but lack the resources necessary to recruit additional staff. For these repositories, many build and maintenance activities are conducted by volunteers. The few full-time programmers that are responsible for platform maintenance are paid using grants and institutional research funding.

By contrast, the other two repositories are larger and have a more structured leadership and governance arrangement in addition to teams of developers. In the case of Repository B, leadership consisted of a distributed network of two co-directors who are supported by a community of leaders, technical experts, researchers, students, and administrative staff. According to one respondent, a third of the team are developers and two-thirds are community support workers who work with project teams to ensure data and CBK sharing standards are upheld. For another repository, demand from a growing community of users necessitated the emergence of two branches: an engineering group responsible for platform development and a Steering Committee focused on future planning, partner management, and prioritization of future tasks.

Procedural Accountability and Preservation Policy: Within TRAC, the description of procedural accountability includes a definition for "designated community." A designated community is "an identified group of potential consumers who should be able to understand a particular set of information" [26, 33]. Each repository studied was born from otherwise unmet needs for replicability, transparency, and harmonization within one or more research communities. Thus, community identification is straightforward but implied and may include anyone with research interests that align with each community. For example, one repository defined its designated community as a community comprised of "anyone excited about open science and machine learning". Otherwise, explicit statements describing the designated communities for each repository were not evident.

Financial Sustainability and Funding: In general, the repositories began either as a grant-funded research or PhD projects. While one repository benefits from large national and institutional sources of funding, the other repositories depend mainly on research and university grants to sustain their operations, making them more financially vulnerable (Table 1). Sustainable business strategies were not yet evident, yet they may be particularly important for these repositories with otherwise limited infrastructure and resources.

Contracts, Licenses, and Liabilities: Contracts are largely absent for these CBK repositories. Instead, they rely on informal community norms or codes of conduct to safeguard the repository from sabotage and uphold intellectual property rights. Two of the four CBK repositories use open-source licensing. Default licenses such as Massachusetts Institute of Technology (MIT) or Creative Commons CC0 Public Domain Dedication licenses, as well as Creative Commons Attribution license (CC-BY) licenses, were used by some repositories, enabling CBK artifact reuse, downstream modification, and further distribution by non-originators of CBKs. One repository employed negotiated control, embargos, limited permissions, and staged sharing or sharing freely within a specified group while offering only limited access to all others.

4.2 Results about Digital Artifact Management

Table 2. Summary of results about three aspects of Digital Artifact Management.

	Repository A	Repository B	Repository C	Repository D
Acquisition	Automatic extraction; users upload	Content from authenticated member-users only	Content from repository users and from select publishers	Content from small group of developers
Ingestion	Active curation and testing	Varies by user and project	Both active curation and testing and acceptance of artifacts "as is"	Active curation and testing
Formal written CBK artifact preservation policy	Not evident	Yes	Yes	Not evident
Access control	No	Yes, by project	No	Yes, some licenses

Under digital artifact management, we examined how the repositories acquired, maintained, and stored content, managed information, and controlled access to their CBK artifacts. A summary of these results is provided in Table 2 above.

Digital Artifact Acquisition: Acquisition strategies varied among the four repositories we examined. As an open-source repository, Repository A allows unmediated user deposit of CBK artifacts. In contrast, as a self-described "community of communities", Repository B gets its content primarily from authenticated member or user contributions. Repository C is somewhat unique, in that it both acquires content directly from contributors, and is part of the publishing pathway for specific journals. Authors deposit models with Repository C and receive an accession number, which is published with the final paper. In contrast to the other repositories, the development team for Repository D, which serves primarily as a distribution platform, is responsible for more than 50% of the artifacts in this repository.

Digital Artifact Ingestion: While the four repositories allow mostly unmediated deposit, we found that there were a variety of intake processes that took place after CBK content has been uploaded to these four repositories.

Repository A, which operates on an open-source model, automatically extracts metadata for storage, and has a two-level review process. In review level one, new CBK artifacts are assessed to be sure that they function correctly; in review level two, CBK artifacts' software code quality is evaluated to determine whether others would be able to work with the resource. This is an open review process. While Repository A currently has a team of 6 core developers, anyone in the community can comment on, discuss, or make contributions to a CBK resource.

Repository C has two different tracks for taking in deposited CBK artifacts. Some of its CBK artifacts are highly curated but others are not. In this case, curated machine learning (ML) models are evaluated for reproducibility against standardized dataset samples and must be able to reproduce at least one result from the related published manuscripts. Curated ML models are also tagged and supplemented with metadata to enhance CBK ML model search and findability; for example, ontologies are used to describe biological pathways, and biomedical concept identifiers are added.

Repository D uses a combination of automated and manual processes for post-deposit intake processing, testing incoming CBK artifact ML models using author-provided tests to be sure that all required fields are complete and that tests pass when conducted by groups other than the original author team. Required information includes a description of the ML model, the URL and path, the framework in which the ML model is implemented, the validation hash for the ML model's code, and a list of ML model dependencies. The ML model may also be tested against a contributor-provided test file to verify expected outputs. The team also verifies that any software packages required to use the ML model are available. Finally, the ML model's license is verified, at which point Repository D will host the materials and make them available for access and download.

Preservation: Maintenance and Archival Storage: Some of the repositories we examined do not have formal preservation or retention policies, but all four CBK repositories generally avoid deleting or removing content from the repository. For example, Repository A may deactivate content but will generally not delete it, as users may still need it as a reference.

Related to long-term CBK artifact preservation, Repository D does not remove content but also does not make any promises with regards to "time to live" (TTL). TTL is limited for pragmatic reasons—e.g., dependencies which are needed for ML models, or systems which manage software packages may themselves become deprecated. Currently, at Repository D, CBK artifacts in the form of ML models are regularly tested through automated processes to be sure they still run, but this is a practice rather than a policy. We found that Repository B does not have a formal service-level agreement for preservation, but it generally follows the CoreTrustSeal requirements although its operators have not yet applied for this seal.

Repository C states that once a resource receives an accession number, it remains in the system for the projected (open-ended) lifespan of the resource. Both Repository B and Repository C receive institutional support, which ensures a different kind of sustainability.

Information and Access Management: Of the four repositories, Repositories A and C are fully open; content in Repository C is released under a Creative Commons (CC) 0 license. Additionally, most of Repository D's content is also fully open, although some resources are restricted by license e.g., no commercial use), and others require authentication via GitHub for access. In contrast, content in Repository B is accessible, but not necessarily open; content contributors are able to set various levels of permission, and embargo periods.

5. DISCUSSION

We reviewed the accessible literature on four CBK repositories and interviewed individuals who are responsible for implementing the policies of these repositories to examine governance structures and strategies for ensuring sustainability and trust. Our aim was to contribute to an understanding of the overall sociotechnical infrastructure required to establish and maintain sustainable CBK repositories within a larger CBK ecosystem. In general, we found that governance was closely aligned with repository maturity, and that initial plans for governance are likely a necessary feature of any group forging a new CBK repository.

For the four actual repositories studied, current approaches to CBK repository organization and governance largely do not align with more formal models described by the TRAC framework. The maturity of the repositories reviewed significantly influences their governance and operational structures. For example, the youngest repository as at the time of review had no governance arrangement while the other three had put in place at varying level of sophistication structures for community development including arrangements to link their operations to industry.

Despite a lack of emphasis on formal governance, primary functions related to CBK ingest, management, access, and archival storage are well-documented across all four platforms. All the repositories reviewed have documented protocols on their websites to guide producers on how to prepare and ingest content into the repository, how the content is managed, accessed, and used. Additionally, in interviews, stakeholders for the four repositories all demonstrated an understanding of their designated communities, even if this knowledge is not explicit on the repositories' websites.

Implications for Repository Management: Our study suggests several key considerations for those engaged in managing CBK repositories. For example, a 'general platform' of CBK repositories operating in a large ecosystem would require a governance infrastructure that establishes standards and caters to the needs of various designated communities. This has downstream implications for the governance of individual repositories that would link into this ecosystem. A resource's designated community determine its policies and practices, and influence its long-term commitment to the retention, preservation, and access to digital information [33]. The starting point for governance of a CBK repository could be articulating a formal mission statement which describes the designated community and specifies linkages between repository mission and its designated community.

Critically, this mission statement could also help to formalize community norms around data description, data sharing, and the responsibilities of content creators and content consumers. This is essential, as the CBK movement brings together stakeholders with different constructs for ethical engagement. While biomedical research is guided by ethical principles of justice, autonomy, beneficence, and non-maleficence, CBK repositories draw on ethics of open science and the principles of FAIRness, ensuring that research products are Findable, Accessible, Interoperable, and Reusable [34]. Additionally, the community ethos of open-source software and code often plays a prominent role in the current governance of CBK repositories. Despite the demonstrated efficacy of this approach within the context of specific repositories, this is not transferrable to clinical settings. A sustainable CBK ecosystem requires a governance framework that ensures rather than assumes the good intentions of the individual members of its community.

Finally, to be trusted by both content contributors and consumers, a CBK repository requires a sustainability strategy that enables it to carry out its mission, despite resourcing challenges. The four repositories we reviewed were funded by grants and their financial procedures and accountability were in accordance with those of their parent organizations. This provides varying levels of security, again tied both to the maturity of the resource, and established external dependencies or relationships. The CBK repositories we interviewed were connected to research enterprises; there are currently limited capabilities to connect to clinical enterprises. There is a need for developing the technical infrastructure and sustainable business models that preserve the interest of their designated communities while harnessing the benefits of private enterprise. While this may depend mainly on individual repositories, the existence of a trusted entity that could certify quality and reliability [20] may serve as a catalyst. Similarly, development guided by stakeholder involvement and CBK maturity and/or deployment models might also be explored.

Implications for Learning Health Systems: The goals of the four CBK repositories studied here suggest the key role of these organizations as central to establishing links between research and practice—a key attribute of learning health systems. While the organizations studied here varied in how long they have been in practice, they all suggested that publishing CBK workflows should be established and formalized. This level of transparency would promote the trustworthiness of CBK repositories as integral parts of the CBK ecosystem, and would promote trust in CBK artifacts, as suggested elsewhere [35]. Further supporting trust and trustworthiness would be clear demonstrations of the feasibility and utility of CBK to deliver knowledge into practice.

Implications for Policy: Establishing standards or best practices for the governance of CBK repositories is a necessary step to ensuring that a robust and sustainable repository infrastructure can meet what will likely be increasing demand for storing and sharing validated CBK. For example, we assume that a large portion of human-readable biomedical knowledge will increasingly be transformed into—or published as—machine readable and executable code that can be deployed to inform patient care. Additionally, we anticipate that learning health systems will have at their disposal applied and adaptive artificial intelligence and machine learning (AI/ML) algorithms to improve health service delivery. CBK repositories may be one mechanism by which validated CBK can be responsibly managed and equitably shared. This will require more mature governance and organizational models than are currently in common practice. It will also require investments that can ensure sustained funding for the development and deployment of a broad range of CBK artifacts. The repositories described in our study provide a public service and/or set of public goods (i.e., CBK) which carry costs that are not covered by open access policies. This is evident in our finding that the current work of CBK repositories is constrained by funding availability for limited range of processes and have emerged from centers that have a heavy academic focus, with limited links to private enterprise. Strengthening this connection is one mechanism for ensuring sustainability. Other business models based on paid membership, for example, might also be explored.

Limitations: This exploratory study has limitations that should be acknowledged. We have examined a narrow slice of the knowledge management landscape, choosing to focus on knowledge management as it relates to medical information and to computable (rather than just digital) knowledge. We also chose to emphasize in our research the issues in knowledge management that relate to the governance as articulated in TRAC, rather than issues related to technology of CBK artifacts. Our research is unique in its empirical approach to understanding current practice related to governance of knowledge repositories in the biomedical field. Consistent with qualitative research studies, our goal is not to generalize broadly but to provide insight into current practices of knowledge repositories in the context of computable biomedical knowledge [36]. Our approach in conducting interviews highlighted important information that is not routinely available on publicly available websites. Future studies should examine technology and additional governance features outside of the scope of the current study. Future studies should also test the generalizability of our findings to other types of repositories in larger knowledge management ecosystem, such as those managed privately or in proprietary EHR systems.

6. CONCLUSION

As the use of CBK artifacts develop and more actors participate in knowledge translation and sharing, more robust organizational infrastructure would be needed to ensure the trustworthiness and effective operations of an ecosystem of multiple platforms. An understanding of how the emerging CBK repositories is governed or managed provides useful inputs into any efforts to develop the sociotechnical infrastructure needed to ensure the effective operation of a federation of digital libraries that curate, store and share actionable computable biomedical knowledge.

AUTHOR CONTRIBUTION

All authors contributed equally to this paper.

P.S. Amara (amaphili@umich.edu) contributed to the design of the research framework, collection of data, and analysis of the data; wrote the introduction, methods, results, discussion, and conclusion sections of the manuscript.

M. Conte (meese@umich.edu) contributed to the design of the research framework, collected the data, contributed to the analysis of the data, writing of the results and discussion sections of the manuscript.

A. Flynn (ajflynn@umich.edu) conceptualized the paper, designed the research framework, contributed to the collection of data; wrote and revised the introduction, methods, results, and discussion sections of the manuscript

J. Platt (jeplatt@umich.edu) proposed the research problem, contributed to the conceptualization of the paper, design of the research framework, collection of data, wrote and revised the introduction, methods, results, discussion, and conclusion sections of the manuscript

M.G. Trinidad (mgracet@umich.edu) contributed to the collection of data, coded, and analyzed the data, and contributed to the writing of the introduction, methods, results, and discussion sections of the manuscript.

REFERENCES

[1] Lynch, C.A.: Institutional repositories: Essential infrastructure for scholarship in the digital age. Portal: Libraries and the Academy 3(2), 327–336 (2003). doi: 10.1353/pla.2003.0039

[2] University of Michigan. Mobilizing computable biomedical knowledge (MCBK) manifesto. Available at: https://mobilizecbk.med.umich.edu/about/manifesto. Accessed 20 October 2021

[3] Boxwala, A.A., et al.: A multi-layered framework for disseminating knowledge for computer-based decision support. Journal of the American Informatics Association 18 Suppl 1(Suppl 1), i132–9 (2011). doi: 10.1136/amiajnl-2011-000334

[4] Mitchell, A.: A NICE perspective on computable biomedical knowledge. BMJ Health Care Informatics 27(2), e100126 (2020). doi: 10.1136/bmjhci-2019-100126

[5] Perrault, R., et al.: Artificial intelligence index 2019 annual report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, December 2019. Available at: https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf. Accessed 20 July 2020.

[6] Chen, P.H.C., Liu, Y., Peng, L.: How to develop machine learning models for healthcare. Nature Materials 18(5), 410–414 (2019). https://doi.org/10.1038/s41563-019-0345-0

[7] Patel, V.L., et al.: The coming of age of artificial intelligence in medicine. Artificial Intelligence in Medicine 46(1), 5–17 (2009)

[8] Alanazi, H.O., Abdullah, A.H., Qureshi, K.N.: A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. Journal of Medical Systems 41(4), 69 (2017). https://doi.org/10.1007/s10916-017-0715-6

[9] Bernstein, S., et al.: A multi-stakeholder roadmap for care transformation—the AHRQ evidence-based care transformation support (ACTS) initiative—S43 Panel 105 (2020). Available at: https://digital.ahrq.gov/sites/default/files/docs/page/amia-ahrq-s43-panel-slides-2020.pdf. Accessed October 15, 2021

[10] Elovic, A., Pourmand, A.: MDCalc Medical calculator app review. Journal of Digital Imaging 32(5), 682–684 (2019). https://doi.org/10.1007/s10278-019-00218-y

[11] Lomotan, E.A., et al.: To share is human! Advancing evidence into practice through a national repository of interoperable clinical decision support. Applied Clinical Informatics 11(01), 112–21 (2020)

[12] Evans, J.P., Dale, D.C.: Fomous, C. Preparing for a consumer-driven genomic age. New England Journal of Medicine 363(12), 1099–103 (2010)

[13] Morris, Z.S., Wooding, S., Grant, J.: The answer is 17 years, what is the question: understanding time lags in translational research. Journal of the Royal Society of Medicine 104(12), 510–520 (2011). https://doi.org/10.1258/jrsm.2011.110180

[14] Flynn, A.J., et al.: Architecture and initial development of a digital library platform for computable knowledge objects for health. Studies in Health Technology and Informatics 235, 496–500 (2017)

[15] Weinstein, J.N.: The health superpowers: Using AI for good. NEJM Catalyst 16, 5(2) (2019)

[16] Sonntag, D., et al.: The clinical data intelligence project. Informatik-Spektrum 39(4), 290–300 (2016). doi: 10.1007/s00287-015-0913-x

[17] Wright, A., et al.: Governance for clinical decision support: Case studies and recommended practices from leading institutions. Journal of the American Medical Informatics Association 18(2), 187–194 (2011). https://doi.org/10.1136/jamia.2009.002030

[18] Kawamanto, K., et al.: A pragmatic guide to establishing clinical decision support governance and addressing decision support fatigue: A case study. AMIA Annual Symposium Proceedings, pp. 624–633 (2018)

[19] Campbell, E.M., et al.: Types of unintended consequences related to computerized provider order entry. Journal of the American Medical Informatics Association 13(5) 547–556 (2006). https://doi.org/10.1197/jamia.M2042

[20] Tutt, A.: An FDA for algorithms. 69 Administrative Law Review 83 (2017). Available at: http://dx.doi.org/10.2139/ssrn.2747994. Accessed July 15 2020

[21] Marcial, L.H., et al.: The imperative for patient-centered clinical decision support. eGEMs (Generating Evidence & Methods to improve patient outcomes) 6(1), 12, 1–8 (2018). https://doi.org/10.5334/egems.259

[22] Tu, S.W., Musen, M.,A.: Modeling data and knowledge in the EON guideline architecture. Studies in Health Technology and Informatics 84(Pt 1), 280–4 (2001)

[23] Shahar, Y., et al.: A framework for distributed, hybrid, multiple-ontology clinical-guideline library, and automated guideline-support tools. Journal of Biomedical Informatics 37(5), 325–344 (2004). https://doi.org/10.1016/j.jbi.2004.07.001

[24] Consultative Committee on Space Data Systems 2011 Audit and Certification of Trustworthy Digital Repositories: recommended practice. Washington DC, Consultative Committee on Space Data Systems, [77pp.] (CCSDS 652.0-M-1). Available at: http://hdl.handle.net/11329/364. Accessed May 20, 2020.

[25] Dale, R.L., Ambacher, B.: Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). The Center for Research Libraries. Available at: https://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf. Accessed May 20 2020.

[26] CRL- The Center for Research Libraries and OCLC Online Computer Library Center, Inc. Trustworthy Repositories Audit & Certification: Criteria and Checklist. 2007. Available at: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac. Accessed May 20, 2020

[27] Lee, C.A.: Open archival information system (OAIS) reference model. Encyclopedia of Library and Information Sciences 3 (2010). Available at: https://ils.unc.edu/callee/p4020-lee.pdf. Accessed October 22, 2021

[28] Cornell University Libraries. Digital preservation management: Implementing short-term strategies for long-term problems (2004). Available at: http://dpworkshop.org/dpm-eng/eng_index.html. Accessed October 10, 2021

[29] Rosenthal, D., et al.: Requirements for digital preservation systems: A bottom-up approach. D-Lib Magazine, 11, 11 (2005). Available at:https://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html. Accessed October 10, 2021

[30] Sharp, C.: Overview of the digital object architecture (DOA). An Internet Society Information Paper. Internet Society (2016, Oct 25). Available at: https://www.internetsociety.org/resources/doc/2016/overview-of-the-digital-object-architecture-doa/. Accessed October 17, 2021

[31] Charmaz, K.: Constructing grounded theory: A practical guide through qualitative analysis. SAGE Publications Ltd, London (2006)

[32] Sebastian, K.: Distinguishing between the strains grounded theory: Classical, interpretive and constructivist. Journal for Social Thought 3(1) (2019). https://ojs.lib.uwo.ca/index.php/jst/article/view/4116

[33] Donaldson, D.R., et al.: Data managers' perspectives on OAIS designated communities and the FAIR principles: Mediation, tools, and conceptual models. Journal of Documentation 76(6) 1261–1277 (2020). https://doi.org/10.1108/JD-10-2019-0204

[34] Wilkinson, M.D., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1), 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[35] Richardson, J.E., et al.: Building and maintaining trust in clinical decision support: Recommendations from the patient-centered CDC learning network. Learning Health Systems 4(2), e10208 (2019). https://doi.org/10.1002/lrh2.10208

[36] Polit, D.F., Beck, C.T.: Generalization in quantitative and qualitative research: Myths and strategies. International Journal of Nursing Studies 1 47(11), 1451–8 (2010). doi: 10.1016/j.ijnurstu.2010.06.004

AUTHOR BIOGRAPHY

Philip S. Amara, MSc, MPH is a Research Area Specialist in the Office of the Provost at the University of Michigan. He was trained in survey methodology at the University of Michigan and in Public Health (Epidemiology) at the University of Nebraska Medical Center. Prior to working in the Office of the Provost, he served as Project Associate Manager in the Department of Learning Health Sciences at the University of Michigan. He also served as Director of Economic Statistics at Statistics Sierra Leone, Country Director at Innovations for Poverty Action and as Monitoring, Evaluation, Accountability and Learning Specialist at the Sierra Leone Ministry of Health and Sanitation. He is a research management practitioner interested in conducting research that generates evidence to improve organizational performance and solve real life problems.

Marisa Conte is the Associate Director for Research and Informatics at the University of Michigan's (U-M) Taubman Health Sciences Library (THL). In this role, she coordinates THL's support for and engagement in research and informatics initiatives across the biomedical research and clinical enterprises. Marisa is also a PhD student in U-M's Department of Learning Health Sciences' Health Infrastructures and Learning Systems program. Her research interests involve computable biomedical knowledge, and the infrastructure required for its maintenance and dissemination. She holds a BA from U-M and an MLIS from Wayne State University.

Allen Flynn

Following his graduation as a Doctor of Pharmacy from the University of Michigan in 1993, Allen Flynn worked as an information system analyst until the dot-com bubble burst. At that point, he began practicing as a hospital pharmacist with information systems responsibilities. Soon Allen was promoted to Coordinator and became involved with several major health IT initiatives. For more than 8 years he held roles of increasing responsibility while gaining expertise in electronic health records and medication system safety. This real-world experience inspired him to try and use information systems to eliminate preventable medication harm. Between 2012 and 2018, he completed a doctorate at the University of Michigan School of Information. Since that time, Allen has taught health informatics and pursued a wide variety of research projects at the intersection of computable knowledge management and medication information systems.

Jodyn Platt, PhD, MPH is an Assistant Professor of Learning Health Sciences trained in medical sociology and health policy at the University of Michigan. Her research focuses on issues at the intersection of informatics and ethics. She is interested in understanding what makes data-driven health trusted and the pathways for earning, achieving, and sustaining trust across stakeholders.

Grace Trinidad completed her PhD in Health Infrastructures and Learning Systems at the University of Michigan Department of Learning Health Sciences. She also holds an MPH from the University of Michigan School of Public Health, and Master of Science in Design in Health from the Taubman College of Architecture and Urban Planning at the University of Michigan. Her research focuses on privacy and trust in all industries, ethics of artificial intelligence and data sharing, and health and data equity.