Request for Information (RFI): Strategic Opportunities and Challenges for the National Library of Medicine, National Institutes of Health

https://rfi.grants.nih.gov/?s=5f15a5e3104800009c001082 SUBMITTED October 15, 2020

Responses are presented through an online form (not a PDF)

Name (limit 255 characters):

Brian S. Alper, MD, MSPH, FAAFP, FAMIA, balper@computablepublishing.com

Name of Organization (limit 255 characters):

COVID-19 Knowledge Accelerator (COKA) at
https://www.gps.health/covid19_knowledge_accelerator.html

Role in Organization (limit 255 characters):

Project Lead; the project is a virtual organization with > 50 participants from > 25 countries

Major opportunities or challenges that have emerged over the last five years and that have implications for the future of NLM in the area of:

a. Science (including clinical health sciences, biomedical science, information science, informatics, data analytics, data science, etc.)

Population health, community health, and individual health are all best advanced by science. Science is best advanced by a large community with an interconnected ecosystem. Effective and efficient community engagement in such an ecosystem requires clear communication. Clear communication in the digital era is supported most efficiently by moving beyond basic digital communication (such as PDF transfer) and achieving computable communication with standards for data exchange, including both the format for the data (covered in the Technology section below) and controlled vocabularies (covered here).

Controlled vocabularies and code systems, such as Medical Subject Headings (MeSH), SNOMED, and RxNorm, support finding, extracting, and transferring knowledge. To make scientific evidence computable and widely accessible, code systems must include the concepts that are commonly communicated and must be accessible to those who are communicating. There are multiple aspects of scientific knowledge that do not have code systems to support common use. We created a Code System Development Protocol (tinyurl.com/CodeSystemDevelopmentProtocol, registered at osf.io/3akjv) to (1) support open collective participation to identify commonly used tools and systems, (2) map the concepts needed to support these systems, (3) map terminologies previously mapped in other ontologies, (4) achieve universal or near-universal agreement on a common code system for functional application, (5) evaluate its implementation, and (6) manage continued updating and maintenance of the code system.

In September 2020, we started efforts to develop a Study Design Code System, a Statistic Type Code System, a Statistic Model Code System, and a Risk of Bias Code System. As of October 15, 2020 we have had 56 people from 26 countries in 6 continents join an Expert Working Group for one or more of these code systems.

Although we are planning to maintain these code systems as a volunteer effort, and we are exploring the National Cancer Institute (NCI) Thesaurus as a virtual space for storing and hosting the code systems, the NLM would be natural home for these code systems which could and should be included in UMLS and interoperate with other NLM terminologies such as MeSH.

b. Technology (including biotechnology, platforms, hardware, software, algorithms, processes, systems, etc.)

Standards for electronic data exchange enable data to be Findable, Accessible, Interoperable, and Re-usable (FAIR). Standards for electronic data exchange are expressed in the form of technology platforms, processes, and systems. A standard for exchange of healthcare data, named Fast Healthcare Interoperability Resources (FHIR), is maintained by Health Level 7 International (HL7), and is becoming widely adopted as the primary standard for healthcare. Rules from ONC and CMMS require that by 2022 all US citizens will have access to all their electronic health data on demand using FHIR.

There is no such standard for electronic data exchange for scientific evidence, namely the results or findings of data analysis including definition of the evidence variables, full expression of the statistics, and interpretation and certainty judgments about the findings. In 2018 we started a project to extend FHIR to express evidence and statistics and have had global participation to the point that today we have FHIR Resources defined for Citation, Evidence, EvidenceReport, EvidenceVariable, OrderedDistribution, and Statistic concepts for data exchange.

The combination of FHIR formats for expression of biomedical science and the code systems to standardize expression of study design, statistics, and certainty of scientific findings will enable extensive and efficient re-use of biomedical knowledge.

Application to ClinicalTrials.gov will add great value to the researchers who submit data to register their studies. Having their data in computable standard form will support their re-use of their own data for subsequent needs such as clinical trial reports for regulatory agencies or creating publications for submission to the peer-reviewed literature.

Any database that NLM supports which provides access to biomedical knowledge, including MEDLINE, PubMed Central, the database of Genotypes and Phenotypes (dbGaP), and others, can be enabled to provide access to a computable representation of the knowledge and support searches and data transformations for a nearly infinite opportunity for current and future technologies to Mobilize Computable Biomedical Knowledge (MCBK).

c. Public health, consumer health, and outreach (including epidemic disease surveillance, culturally competent engagement, optimizing the experience of resource users, etc.)

Application of standards to make data more FAIR for both data entry and data use will go a long way towards optimizing the experience of resource users. We are developing a Feasibility and Usability

Testing protocol to support this optimization. (draft in progress at https://docs.google.com/document/d/1SUWtErihVBks3XUO6L1fW5WPD3RcP6_j)

d. Library functions (including collection development, access, preservation, indexing, library metadata, service agreements with other libraries, etc.)

The NLM is not just a home for substantial collections but could be considered a leader in library function, interoperability, re-use, and sharing. A common framework for referring to entries in collections will enable expansion of functions throughout all libraries. If the "common framework" is limited to NLM-specific concepts, the NLM networks will be limited. If a broader "common framework" can be reached with more communities serving library functions, there will be far greater impact.

The FHIR Citation Resource (http://build.fhir.org/citation.html) has about 100 elements and conveys all the data found in a MEDLINE citation for a journal article as well as data needed for other citation considerations, such as citing articles published in databases or books or citing computable resources that do not have structures similar to journal articles.

The FHIR Citation Resource can facilitate interoperability and data exchange across citation repositories, such as PubMed and COVID-19 Open Research Dataset (CORD-19). We have produced a detailed mapping of MEDLINE elements (XML, text and RIS formats) to the FHIR Citation elements to facilitate expression of MEDLINE citations in FHIR format.

e. Modes of scholarly communication (including researchers' use of social media, preprints, living papers, changes in the roles and practices of publishers, data-driven approaches to studying historical medical texts, images, and datasets, etc.)

Important considerations in scholarly communication include attribution of scholarly contributions. Specifically there are needs to (1) document multiple types of contributions other than "authoring", (2) distinguish whether contributors are or are not "authors", (3) apply contributorship attribution to conventionally published articles, and (4) apply contributorship attribution to knowledge artifacts that are not conventionally published articles.

The FHIR Citation Resource (http://build.fhir.org/citation.html) includes a contributorship element that covers all the needs found in author entries in MEDLINE entries today but also includes additional elements. The contributorship.entry.notAnAuthor element allows simple toggle to distinguish author from non-author contributors. The contributorship.entry.correspondingAuthor element allows a simple toggle to identify the preferred contact point among the contributors.

The contributorship.entry.contribution element allows coding any number of contribution types and we currently suggest a code system based on the CRediT Taxonomy at https://jats4r.org/credit-taxonomy.

The contributorship.summary element allows data exchange for complete author lists or contributorship statements as an alternative or supplement to individual entry listing. The FHIR Citation Resource itself can be applied to both conventionally published articles and other knowledge artifacts.

The CRediT Taxonomy is a substantial advance over unstructured contributorship statements but it is not sufficient in its current form to account for formal reviewer and editor roles in the larger universe of

scholarly contributions. Also, the CRediT Taxonomy alone is not sufficient to account for contributions to components of multi-component scholarly productions, such as systematic reviews where CRediT Taxonomy items can be applied to the search effort, to the evidence selection effort, to the data extraction effort, to the critical appraisal effort, etc.

We would welcome an opportunity to combine efforts to "represent the roles typically played by contributors to scientific scholarly output" (the purpose of CRediT stated at https://casrai.org/credit/) and efforts to provide multifaceted global engagement (as described in the Code System Development Protocol noted previously) to provide a code system for contributions to scientific scholarly output. NLM engagement would be a substantial opportunity to provide a home for such efforts.

f. Perspectives, practices, and policies (including those related to open science, the need for diversity, equity, and inclusion in research, algorithmic bias, expectations of reproducibility of research, etc.)

...left blank...

g. Workforce needs (including data science competencies, effective strategies for recruitment and retention of underrepresented minorities, opportunities for training and continuing education for middle- and late-career researchers and librarians, etc.)

...left blank...

Major opportunities or challenges that have emerged in the last five years and that have implications for the future of NLM in other areas or areas not well captured above.

The rapid adoption of FHIR for healthcare data, and Maturity Model Level 1 development of multiple FHIR resources to support computable expression of scientific evidence, provides many wide-reaching opportunities for NLM.

Opportunities or challenges on the horizon over the next five years that fall within the purview of the NLM's mission.

Many have hopes for artificial intelligence and machine learning to accelerate and advance our knowledge application. Although this appears to work well in many areas of descriptive analytics and predictive analytics there is currently a large gap between desires and realities for prescriptive analytics. One of the biggest barriers for prescriptive analytics attempting to use natural language processing for recognition of the "truth" (or reference standard or training set for machine learning) is that our natural language for expression of effects of healthcare interventions is imprecise and ambiguous. Thus any machine processing compounds the problem rather than provides reliable prescriptive analytics. Transformation of our scientific communication to precise, unambiguous, codable concepts in standard formats for data expression (ie, making science machine-interpretable) will be the key to empowering artificial intelligence for prescriptive analytics.