Melissa Haendel <https://orcid.org/0000-0001-9114-8737>

Anita Walden <https://orcid.org/0000-0002-3327-7423>

James Eddy <https://orcid.org/0000-0001-9758-0176>

Samuel Volchenboum <https://orcid.org/0000-0001-9863-851X>

<https://grants.nih.gov/grants/guide/notice-files/NOT-LM-21-008.html>

Original RFI: <https://grants.nih.gov/grants/guide/notice-files/NOT-LM-21-005.html>

**Information Requested**

NIH is requesting public comment on the use of CDEs, particularly in the context of COVID-19 research, including opportunities for advancing research with CDEs, challenges to adopting CDEs, and guidance or tools that could facilitate use of CDEs. These comments will be used to inform NIH’s continuing development of guidance of CDE use for COVID-related research and assist in the planning for adequate funding of CDE efforts through research awards and contracts.

Response to this RFI is voluntary. Respondents are free to address any or all topics listed below, as well as other relevant topics, for NIH’s consideration.

1. **Current use of Common Data Elements in research projects.** NIH seeks broad input on how researchers are currently using CDEs, regardless of whether research is related to COVID-19.
   1. Explain how you identify, select, deploy, and encourage use of CDEs in your research
   2. Describe the benefits to you of using CDEs in your research.
   3. Identify existing incentives or resources that have facilitated your use of CDEs.
   4. Describe COVID-19 research activities you have engaged in and how you have used CDEs to collect data.

*Draft response:*

|  |
| --- |
| **Our groups**  The collective groups represented by this response include the NCI funded Center for Cancer Data Harmonization, the Pediatric Data Commons, the NCATS funded Center for Data to Health, and other projects. Our teams utilize CDEs extensively in our prospective studies as well as in our data harmonization efforts. We participate in many SDOs such as HL7, ISO, CDISC, GA4GH, WHO, etc.  **Our use of CDEs**  Through the CD2H program and in collaboration with the NCI, we developed and led the Metadata Automation DREAM Challenge to address the time-consuming task of retrospective metadata harmonization. Using structured biomedical data files, challenge participants developed containerized tools to automate annotation of metadata fields and values, using available research data annotations (caDSR CDEs) as well as established terminologies and ontologies. Tool-annotated CDEs and permissible values were evaluated against manually curated gold standard metadata to assess performance. In the context of the National COVID Cohort Collaborative (N3C), we have leveraged the computationally accessible codesets within the caDSR to help inform the Common Data Model mapping for content coming into the EHR via survey instruments or other CDE contexts.  We have also participated in pediatric COVID-19 CDE development. In other work funded by an R03, we have harmonized prospective study data to NIH CDEs and shared them in NIH repositories.  Fundamentally, the lack of computational encoding of CDEs has limited the ability to collect data in a born interoperable fashion. While many CDE registries exist, only one has a foundation in computational encoding - the caDSR. We have therefore taken great advantage of the RDF, API, and semantic encoding of the codesets to support our data harmonization activities as well as to choose appropriate CDEs for our use. We have used the LinkML modeling language for the development of harmonized data models, this provides us an opportunity to leverage and incorporate the semantic elements that were previously “bolted on” to other models via the caDSR. LinkML allows modelers to simultaneously describe what an instance of a data record must look like and what the components in the model actually mean. The [ISO/IEC 11179 standard](https://www.iso.org/standard/50340.html), a metadata registries metamodel of the International Organization for Standardization (ISO)’s technical committee on data management and interchange, evolved, in part, around the SQL standards developers’ need to add semantics to SQL data structures -- to be able to determine when columns from tables in different databases describe the same thing and to identify how one might go about transforming information from a source table/column to its analogous target -- i.e., *data model harmonization*. The LinkML foundation of the model harmonization efforts puts us into a position to leverage and *utilize* the background 11179 work. It enables:   * Incorporation of existing 11179 content in the NCI caDSR: we are leveraging the RDF representation of the caDSR content to incorporate the existing semantic annotation work in both the source (e.g., PDC, CDH) and target (CCD-H) models. * Eventual export of new and enhanced semantic annotation back to the caDSR in the form of RDF or other target format. * Documentation *and* validation of extant and proposed “value sets” (which we have chosen to call “code sets” for the sake of clarity) along with their associated permissible values.   **Assess Landscape for Existing CDEs**  For most project efforts, we review NCI’s CDE Browser, CDISC, HL7 and other existing sources for well vetted CDEs. The groups mentioned typically have a robust development and vetting process for their CDEs that include multiple stakeholders such as professional societies, SDOs government agencies, Pharmaceutical organizations, academic medical centers, clinician groups, informaticists, vendors and others. These diverse groups meet to develop the standards, use an open process for broad development and vetting and follow a criteria for addressing comments. If the above CDEs do not meet the needs they are modified for the specific purpose. There is a set of common data elements that typically can meet most needs. Many industry companies have standard libraries with governance that allow for consistent use of CDEs primarily based on CDISC standards but they may also include HL7 and other domain specific data elements.  **Develop if CDEs don’t exist**  If the CDEs don’t exist, we establish CDEs with a cross section of healthcare, research, registry and other stakeholders. This is only if there is time to go through a SDO well-vetted process. If there is not enough time to create standard CDEs where they don’t exist, then we have used what was common and created new data elements for the specific project. It typically takes 9 months to create a standard with multiple stakeholders (includes development, vetting, formatting). We have demonstrated cost saving using CDEs and cost saving using Standard CDEs over multiple studies.  **Demonstrate Cost Savings**  The cost savings is for downstream projects. The first project may cost more initially but the savings are increased with each additional use of the CDEs. Many large industry organizations have CDE libraries with governance that reduce the number of custom CDEs, many based on CDISC. Some Academic organizations have also used standard terminologies, HL7 and other CDEs to reduce their research costs and improve analytics. |

1. **Challenges to adopting CDEs.** NIH seeks broad input on barriers to adopting CDEs in the context of both research activities generally and COVID-19 research specifically.
   1. Describe the major barriers you have encountered in using CDEs.
   2. Provide specific examples of challenges you have encountered in identifying and/or using CDEs in your research.
   3. Describe why you have or have not chosen to use CDEs in your research.

*Draft response:*

|  |
| --- |
| Requirements take longer than most people expect and need standardized processes. We encourage the thoughtful, community-based requirements driven creation and alignment of CDEs prior to new research wherever this can be accomplished. Governance, decision making, and community followup for CDE evolution are critical for prospective interoperability.  **Finding CDEs is hard.**  Many CDE registries exist, and most seem to include a significant amount of content from each other. PhenX, NLM CDE browser, caDSR, the Value Set Authority, HL7, CDISC - all have overlapping content. How is this kept synchronized? What content is actually synchronized? What is the versioning, provenance and attribution? This means that when a data element is changed due to new science or another reason you want to track the version that was used and where it was used (protocol, report, dataset, case report form…). This complexity and variations of rigor makes use of CDEs daunting and promotes a culture of reinvention. Further, some of the content is proprietary (not available to the public or there is a cost associated with using the content), making mashup and redistribution unclear and access not always certain.  **Computational representation.**  One of the biggest challenges in the CDE landscape is that they are often created by communities without computational expertise. This means taking a data element and definition with a standard set of permissible values and placing it in a format that can be used by a computer. Many are in word, excel or PDF documents which can not be used by a computer. While good conceptual alignment can be achieved by these communities (such as in the case of PhenX community development and governance processes, which are very good), there is rarely a computational encoding or validation. This is usually attempted after the fact, and is required for downstream data harmonization and conversions. Downstream data harmonization and data conversions can be difficult and the effort required may outstrip available resources (e.g. R03). To date, caDSR is the only computationally encoded resource available.  **Costs, Time and Training.**  Over the years when developing and using CDEs the challenge is the implementation costs and tooling. The systems used in healthcare and research should be improved to incorporate standards easily. Smaller organizations don’t have the funding to implement standards even if there is a downstream savings. Standards implementation can significantly delay startup and increase quality issues due to changes to systems and workflow. Larger organizations can absorb these costs until they see the savings usually after the 2nd and 3rd study/database (based on our analysis).  **Tooling** will help with standards implementation, enhancements, and maintenance. The right type of **Training** is also needed to learn the standard, understand how it can impact the studies, systems, and workflow. Again, smaller organizations usually don’t have the resources to learn or pay for training. Patience...is needed to implement even the best written standard!!! We often don’t have patience to work through the complexities. |

1. **CDE guidance, tools, infrastructure, and incentives.** NIH seeks broad input on what guidance, tools, and incentives would facilitate wider uptake of CDEs.
   1. Suggest specific guidance or information that would enhance your understanding and use of CDEs.
   2. Describe how the [NIH CDE Repository](https://cde.nlm.nih.gov/home) could be enhanced to facilitate discovery and use of relevant CDEs.
   3. Describe resources that would enable use of CDEs, such toolkits for including CDEs in studies or tools that support researchers’ mapping existing data to CDEs.
   4. Describe incentives that would encourage and increase use of CDEs.
   5. Provide specific examples of questions you have about how to identify, select, deploy and use CDEs.

*Draft response:*

|  |
| --- |
| Three key things would improve the community understanding and use of CDEs. First, there should be tutorials and tools that make it easy to reuse them. We have different resources that each do a different thing best: the NLM CDE registry has a nice string search and display, but is not computationally encoded or semantically interoperable. The PhenX registry has user-friendly structure to navigate related CDEs and collections and a terrific community creation, validation, and governance structure, but again is not computationally encoded (though there are some tags). The caDSR is the most comprehensive and is computationally far and away the most robust for actual use in software systems, but is very challenging to search and the mechanisms by which content gets added and encoded are not clear. The Value Set Authority has computationally encoded value sets but is not well integrated with instruments or other CDE resources.  Another key issue is a lack of community coordination and transparency around how and where to create the CDEs. There are no fewer than 8 different NIH groups making COVID-related CDEs, but because this work is not being created transparently, this promotes a culture of reinvention as the community cannot wait. It is not clear on the differences between the long standing mature CDE Browser and the VSAC and the governance across them. When to use one versus the other and the process for submission. Only those who have been deeply involved in developing standards know the landscape but most have questions such as:   * Does the NLM CDE repository also adhere to ISO 11179 like the NCI CDE Browser? Are there other identifier or format standards, conventions that are suggested / required? * What are the metrics for inclusion? While they require computability? Will that align with the model realized in caDSR? * Is there any publicly-available information about the trans-NIH endorsed CDE process? How does that happen? Related to this: what is the process to propose candidate CDEs? What requirements must be met and then how are they asking you to select and promote selected members onto the platform? * What (if any) relationships are there between the NIH CDE Repository & the VSAC? Is NIH planning on interdigitating these 2 platforms, or will they remain separate facilities?   Utilizing a central Standards Development Organization (SDO) to create computable CDEs in partnership with Clinical and Domain Subject Matter experts, Informaticist, Implementers, Vendors and Tool developers will help with the entire process. An accredited SDO process will ensure a formal process for development, approval and maintenance which can reduce one off development, improve cross-functional stakeholder development and robust guidelines for contribution, vetting and transparency. This has been done before with cardiology, infectious disease and emergency services standards. The challenge is that organizations create their own without due diligence to look for existing standards and are not open to the option of using and modifying what exists.  For tooling, in an ideal world you would have templates, such as in CEDAR or REDCap (not open source, so harder to contribute) that would call upon the rich encodings of the CDEs within caDSR. LinkML or other common modeling language would support creation and alignment of a CDE schema. There would be perfect synchronization of the front ends of NLM and PhenX so that users can find what they are looking for. The codesets and their enumerations would have resolvable URIs, and be associated with identified data model fields. Improving the use of standard schemas rather than allowing the current proliferation (made worse by FHIR Implementation Guides (IGs) with no agreement a priori on the modeling and semantics) would exist that are associated with the codesets, making the overarching semantics of how the CDEs fit into the data management, validation, and harmonization strategies more easily obtained.  Incentives:  **Resources** Funding to train, implement, maintain and design tooling. Also, incentive to manage changes. Every couple of years a new CDE list or standard comes out which is costly to implement. Organizations that manage changes to their CDEs have better uptake because it reduces the burden on those implementing the standard. Many EHR Vendors are skeptical on the “Next Best” CDE group because when funding changes the CDEs will change too.  **Common CDEs** across all Domains. Use a Common set of CDEs that everyone can use. Revisiting Race, Ethnicity, Gender with every CDE group is not a good use of time. Reuse what is available and extend it for the minor changes.  **Authoritative Sources**. Trace authority. Who created the CDEs is important. Indicate if they are developed by a professional society, based on national or international guidelines, developed by a cross section of stakeholders vetted through a standards development organization such as HL7, IHE, CDISC… CDEs that have some sort of authority are trusted rather than from a small group that organized.  **Carrot and Stick**  Most groups will not use standards or Common CDEs unless they have too. Included the use of the CDEs in FOAs. For example CDISC was used by early adopters but not by the masses until FDA required it. Partner with Vendors. They want to know if there is a large international group and regulatory weight behind the CDEs (ONC, FDA, NIH, HL7, CDISC...).  **Show me the money**  Showing actual financial figures that demonstrate savings to the people who are implementing the standards. When we showed the large funding agencies and pharma companies the costs of not using Standard CDEs vs using them that convinced them to invest.  Support for investment! |