NIH Master Reference Dictionary Proposal: Rationale, Requirements, and Cost Estimate

Report from the Master Reference Dictionary Working Group (MRD WG)

Presented to the NIH Common Data Elements Task Force (CDE TF) of the NIH Scientific Data Council (SDC)

January 7, 2019

Contents

NIH Master Reference Dictionary Proposal: Rationale, Requirements, and Cost Estimate 1
Executive Summary 4
Background and Contributors 6
Master Reference Dictionary Working Group 6
Data Elements Dictionary Infrastructure Working Group 6
Rationale 7
Functional Requirements (Core Services) 11
Discovery, Search (Basic & Advanced), Retrieval, and Download 11
Metadata Driven Software Development 12
Vocabulary & Terminology Services 12
Registration, Submission, Governance, Audit, Curation, Harmonization, and Maintenance 12
Reporting and Administration 14
Cost Estimate for NIH Master Reference Dictionary (a library of common data elements) 15
Table 1: INDEPENDENT GOVERNMENT COST ESTIMATE (IGCE) for Base Period (Year 1) 16
Table 2: INDEPENDENT GOVERNMENT COST ESTIMATE (IGCE) for Years 2-4 17
Appendix A: Definitions 18
Appendix B: GSA Service Rates Used to Develop Cost Estimate 20
Appendix C: Figures from Infrastructure Working Group White Paper 21
Matrix showing characteristics of CDE development and usage 21
Current Common Data Elements Environment 21
Proposed NIH Master Reference Dictionary to serve as the core component for proposed NIH CDE infrastructure (Phase 1) 22
Appendix D: Benefits of and Challenges to the Use of CDEs 23
Appendix E: Executive Summary from 2015 NIH CDE Workshop 24

Appendix F: Infrastructure and Tools Group "Birds of a Feather" Summaries from the 2015 NIH CDE Workshop 26
Appendix G: Summary of Systems & Standards Presentations 30
Appendix H: High Quality, Reusable Data for Biomedical Research: Introduction to Clinical Common Data Elements (CDEs) 32
What is a Common Data Element (CDE)? 32
When should CDEs be used? 32
Who can CDEs help? 33
How can CDEs help us get high quality, reusable data for biomedical research? 33
What are the challenges in developing and selecting CDEs? 34
Overcoming challenges. 34
Building upon best practices 35
References 35
Appendix I: Common Data Elements: Promising Practices and Lessons Learned from NIH Experience 36
Lesson 1: Don't reinvent the wheel – Determine what information you wish to gather, and whether data elements of interest already exist. 37
Lesson 2: Don't work alone - coordinate with other ongoing efforts 38
Lesson 3: Engage the full range of needed expertise - domain experts, informatics experts, community experts, and other relevant experts as the content changes. 38
Lesson 4: Use data elements that are scientifically validated, whenever possible. 39
Lesson 5: Use data elements that conform to existing data standards and are freely available, whenever possible. 39
Lesson 6: Plan ahead to update CDEs over time and keep track of all versions. 40
Lesson 7: Gain support of the user community. 41
Lesson 8: Develop a communications strategy 41
Lesson 9: Develop policies to promote the use of CDEs 42
Lesson 10: Monitor compliance and uptake/use of CDEs 42
Lesson 11: Follow FAIR Data Principles in distributing CDEs 43
Lesson 12: Harmonize Across Paradigms 43
Appendix J: Policy Language to Encourage Use of NIH CDEs 44
NIH Template Language 44
IC-Specific Language 44
Request for Applications: 44
Notice of Grant Award: 46

Guide Notices: 46
Other Approaches 47
Appendix K: Examples of IC and non-NIH Practices for Monitoring Investigator Compliance with CDE and Data Sharing Policies 48
FITBIR 48
NINDS 48
NDAR 48

Executive Summary

In the context of clinical research, broad adoption of common data elements (CDEs) can reduce burden for researchers during both study design and post-study analysis, and increase comparability across studies. In addition to streamlining project implementation and analysis, aligning questions and their responses can help facilitate cross-study collaborative research analysis and comparison. When data are collected in a pre-defined way, it lessens the burden on analysts downstream by reducing the need to map or transform the content of data.

NLM's 2017-2027 Strategic Plan identified common data elements as a means by which to facilitate interoperability of data. NLM hosts the NIH CDE Repository which began development in 2012. Since then, all support for the NIH CDE Repository has been provided by NLM, through initial and recent funding from the Office of the Secretary Patient-Centered Outcomes Research Trust Fund for Data Infrastructure (PCORTF), as well as NLM appropriated funds. Launched in 2015, the NIH CDE Repository provides electronic formats of data elements and forms to be used for data collection, sharing and reuse, for clinical research and other purposes.

The NIH CDE Repository is mentioned in the NIH Strategic Plan for Data Science as a possible means by which NIH might track progress towards the goal to "Promote Modernization of the Data-Resources Ecosystem." This strategic plan recognizes CDEs as data standards that "help improve accuracy, consistency, and interoperability among data sets within various areas of health and disease research". NIH Institutes, Centers, and Offices (ICOs) are at different stages of interest in, experience with, development of, recommendations about, and use of CDEs.

To enable CDE data to be findable, accessible, interoperable and reusable (FAIR), the Master Reference Dictionary Working Group to the NIH Common Data Elements Task Force recommends that NIH allocate funding resources for development and maintenance of the NIH Master Reference Dictionary (MRD), related curation and coordination support services, and customer service/technical assistance functions. This high-level trans-NIH investment and commitment will help ensure that the MRD is a sustainable resource that will continue to benefit the biomedical research community.

A fundamental goal of the NIH is to generate good and reusable data and to make data sharing meaningful. The MRD would help harmonize data elements and forms to that end. The Working Group recommends engaging subject matter experts to review and, where appropriate, help to standardize and harmonize NIH CDE content. For this effort to be successful, it is critical to establish a rigorous and well-defined governance process with clear scope and authority to set criteria for reviewing CDE content submitted by ICOs and promote the use of NIH-wide preferred CDEs. While respecting and enabling ICO-level subject matter expertise and focus, the MRD would also enable the NIH community to leverage shared technical infrastructure and other complementary shared resources.

The MRD would be an NIH-wide reference library for key common data elements and forms, supported by tooling that can work interoperably with the diverse data repositories and metadata repositories also funded by NIH (by specific ICOs). The MRD would be a "next generation" of the existing NIH CDE Repository currently developed and hosted by the NLM. The MRD could be created by developing additional features on top of the existing CDE Repository software platform, or by starting with and customizing a vendor-developed metadata registry software product.

The following five categories of functional requirements are critical for phase 1 of the NIH MRD:

- Discovery, Search (Basic & Advanced), Retrieval, and Download
- Metadata Driven Software Development
- Vocabulary & Terminology Services
- Registration, Submission, Governance, Audit, Curation, Harmonization, and Maintenance
- Reporting and Administration

Detailed lists of the critical tasks and services identified for each of these categories are included in the Functional Requirements section of this proposal.

Importantly, adoption of CDEs requires more than just a technical software infrastructure solution. To complement the technical capacity of the MRD, the Working Group also emphasizes the need for ongoing work, led by the NIH CDE Task Force, to:

- Coordinate and develop educational and training resources about how to use CDEs in general, and how to use the NIH MRD, customized as appropriate for numerous stakeholders including investigators, grantees, program/project/scientific review officers, and CDE curators;
- Develop incentives for research initiatives to adopt CDEs (e.g. more widespread use of FOA language that strongly encourages or requires use of CDEs); and
- Determine key metrics and methods for tracking development, adoption, and use of the MRD and of CDEs through the workflow of research: at the study design level, during data collection, publishing, and for secondary use in data sharing platforms.

Background and Contributors

In 2015, the NIH Common Data Elements (CDE) Task Force charged the Data Elements Dictionary Infrastructure Working Group to develop a white paper as a response to NIH Common Data Elements (CDE) Workshop Group 3 activities to "coordinate and integrate NIH technical and software infrastructure(s) for supporting CDE development, discoverability, and access". The Infrastructure WG, with representatives of different ICOs with experience in CDEs, reviewed current infrastructure, discussed strengths and weakness of current systems, and developed representative use case scenarios to serve needs of all the stakeholders, including investigators, curators, informatics specialists, and NIH staff. Portions of the white paper, presented to the CDE Task Force in February 2018, are incorporated into this report. Among the recommendations of the white paper was the development of an NIH Master Reference Dictionary (MRD).

The Task Force then charged the MRD Working Group to develop a proposal and estimate the cost to create the MRD. The MRD is envisioned as an NIH resource with shared infrastructure, investment, and governance. The MRD would be made usable for all NIH ICOs directly through web interface and APIs. It is intended to coordinate and integrate NIH technical and software infrastructure(s) for supporting CDE development, discoverability, and accessibility. The MRD WG identified and prioritized critical tasks and services as functional requirements of the NIH MRD for phases 1 and 2, then estimated the cost based on experiences of several Institutes, Centers, and Offices (ICOs) with CDE and data registry initiatives, including input from ongoing NCI contract exercises for modernization of the NCI Metadata Registry.

Master Reference Dictionary Working Group

Chair: Carolina Mendoza-Puccini **(NINDS)**
Contributors (listed alphabetically by IC then surname)

- Matt McAuliffe, Vivek Navale, Olga Vovk (**CIT**)
- Denise Warzel (**NCI**)
- Kerry Goetz, Santa Tumminia (**NEI**)
- Natalie Pino, Erin Ramos, Heidi Sofia (**NHGRI**)
- Lucy Hsu, Ellen Werner (**NHLBI**)
- Andrew Weitz (**NIBIB**)
- Sharon Lawlor, Jenna Norton, Kenneth Wilkins (**NIDDK**)
- Greg Farber (**NIMH**)
- Xinzhi Zhang (**NIMHD**)
- Joan Austin, Donna (DJ) McCloskey (**NINR**)
- Liz Amos, Rebecca Goodwin (*CDE TF Executive Secretary*), Vojtech Huser, Lisa Lang, Anna Ripple (**NLM**)

The Master Reference Dictionary Working Group was charged to review and expand upon the recommendations of the Dictionary Infrastructure Working Group to develop an NIH Master Reference Dictionary. The MRD WG was asked to define the scope and core functional requirements of such a resource, develop a preliminary cost estimate, and gauge whether NIH and ICOs might commit to supporting and utilizing the resource.

Data Elements Dictionary Infrastructure Working Group

Chair: Matthew McAuliffe, **CIT**
Contributors (listed alphabetically by surname with affiliation): Liz Amos, **NLM**, Ian Fore, **NCI**, Tsega Gebremicael, **CIT-Sapient**, Kerry Goetz, **NEI**, Vojtech Huser, **NLM**, Lisa Lang, **NLM**, Donna Jo McCloskey, **NINR**, Vivek Navale, **CIT**, Rachael Roan, **NLM**, Yaffa Rubinstein, **NLM**, Denise Warzel, **NCI**.

The Dictionary Infrastructure Working Group was charged to develop a white paper as a response to NIH Common Data Elements (CDE) Workshop Group 3 activities to Coordinate and integrate NIH technical and software infrastructure(s) for supporting CDE development, discoverability, and accessibility. Portions of the white paper, presented to the CDE Task Force in February 2018, are incorporated into this report.

Rationale

In the context of clinical research, broad adoption of common data elements (CDEs) can reduce burden for researchers during both study design and post-study analysis, and increase comparability across studies. In addition to streamlining project implementation and analysis, aligning questions and their responses can help facilitate cross-study collaborative research analysis and comparison. When data are collected in a pre-defined way, it lessens the burden on analysts downstream by reducing the need to map or transform the content of data.

For example, if researchers asked about a person's marital status in Study A with answer choices of "married or single" and tried to combine this data with Study B which allowed subjects to answer: "married; single; divorced; partnered, widowed, or chose not to answer", harmonization would be difficult if not impossible. This is a simplified example of the types of challenges that are created when common data elements are not established before data for a study is collected.[1]

CDEs are also crucial in enabling the linkage and reuse of research and routine healthcare data recorded in Electronic Health Records, as demonstrated by ongoing NIH-FDA collaborations for access to and use of registry data for women's uro-gynecological conditions. This link is realized through annotation of CDEs (limited to those that partially or fully overlap with routine healthcare data collection needs) with terminologies required to support the interoperability efforts of the Office of the National Coordinator for Health Information Technology (ONC).

The National Library of Medicine (NLM) hosts the NIH CDE Repository which began development in 2012. Since then, all support for the NIH CDE Repository has been provided by NLM, through initial and recent funding from the Office of the Secretary Patient-Centered Outcomes Research Trust Fund for Data Infrastructure (PCORTF), as well as NLM appropriated funds. Launched in 2015, the NIH CDE Repository provides electronic formats of data elements and forms to be used for data collection, sharing and reuse, for clinical research and other purposes.

[1] Corwin, E. J., Moore, S. M., Plotsky, A., Heitkemper, M. M., Dorsey, S. G., Waldrop-Valverde, D., Bailey, D. E., Docherty, S. L., Whitney, J. D., Musil, C. M., Dougherty, C. M., McCloskey, D. J., Austin, J. K. and Grady, P. A. (2017), Feasibility of Combining Common Data Elements Across Studies to Test a Hypothesis. Journal of Nursing Scholarship, 49: 249–258. doi:10.1111/jnu.12287

NLM's 2017-2027 Strategic Plan identified Common Data Elements as a means by which to facilitate interoperability of data.[2]

The NIH CDE Repository is mentioned in the NIH Strategic Plan for Data Science as a possible means by which NIH might track progress towards the goal to "Promote Modernization of the Data-Resources Ecosystem."[3] This strategic plan recognizes CDEs as data standards that "help improve accuracy, consistency, and interoperability among data sets within various areas of health and disease research". NIH Institutes, Centers, and Offices (ICOs) are at different stages of interest in, experience with, development of, recommendations about, and use of CDEs.

This attention to CDEs and the acknowledgement of the need for tracking metrics for data sharing suggest that NIH Leadership is willing to collaborate with the NIH CDE Task Force, NLM, and other ICOs to provide the governance measures necessary to encourage increased use of CDEs. The NIH CDE Task Force believes that those efforts would be supported through the development of the Master Reference Dictionary (MRD) to include IT infrastructure for cataloging and tracking CDE usage.

Key Characteristics of Proposed NIH Metadata Reference Dictionary

- Make CDEs "FAIR" – findable, accessible, interoperable, reusable.
- Centralized NIH technical and software infrastructure(s) for supporting CDE development, discoverability, and access. Enable ICOs to share CDE collections and allow stakeholders to find, reuse, and suggest new CDEs.
- Detailed architecture adhering to software best practices.
- Common/core definitions of attributes of CDEs required for system interoperability, including unique identifiers and information required for version control of CDEs and forms.
- Platform to enable harmonization.
- Common formats for CDE import/export.
- Common application programming interface (API) for accessing NIH CDE-related resources.
- Rigorous and well-defined governance process with clear scope and authority to set criteria for reviewing CDE content submitted by ICOs and

Figure 1: Key Characteristics of Proposed NIH Metadata Reference Dictionary

The MRD Working Group believes that consistent with these strategic plans, the time is right to coordinate and integrate NIH technical and software infrastructure(s) for supporting CDE development, discoverability, access, and utilization. To accomplish this, we propose to develop and support the NIH MRD as an NIH resource with shared infrastructure, investment, and governance. The MRD would be made usable across NIH via centralized web interface and APIs. The improved design should support harmonization of data elements, with auditing capabilities, to enable discovery, search and retrieval of all forms that use CDEs.

The working group recommends that NIH allocate funding resources for development and maintenance of the MRD and related curation and coordination support services, to enable CDE data to be findable, accessible, interoperable and reusable. This high-level trans-NIH investment and commitment will help

[2] NLM Strategic Plan 2017-2027 (https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.html) accessed January 3, 2019

[3] NIH Strategic Plan for Data Science (https://datascience.nih.gov/strategicplan) accessed January 3, 2019

ensure that the MRD is a sustainable resource that will continue to benefit the biomedical research community.

As a proposed integrated and unified NIH CDE infrastructure system design (Appendix C), the MRD would be very useful for increasing research data integration, supporting FAIR initiatives (Findable, Accessible, Interoperable, and Reusable) for accelerating knowledge discovery through data sharing and reuse by helping to harmonize data elements and forms.

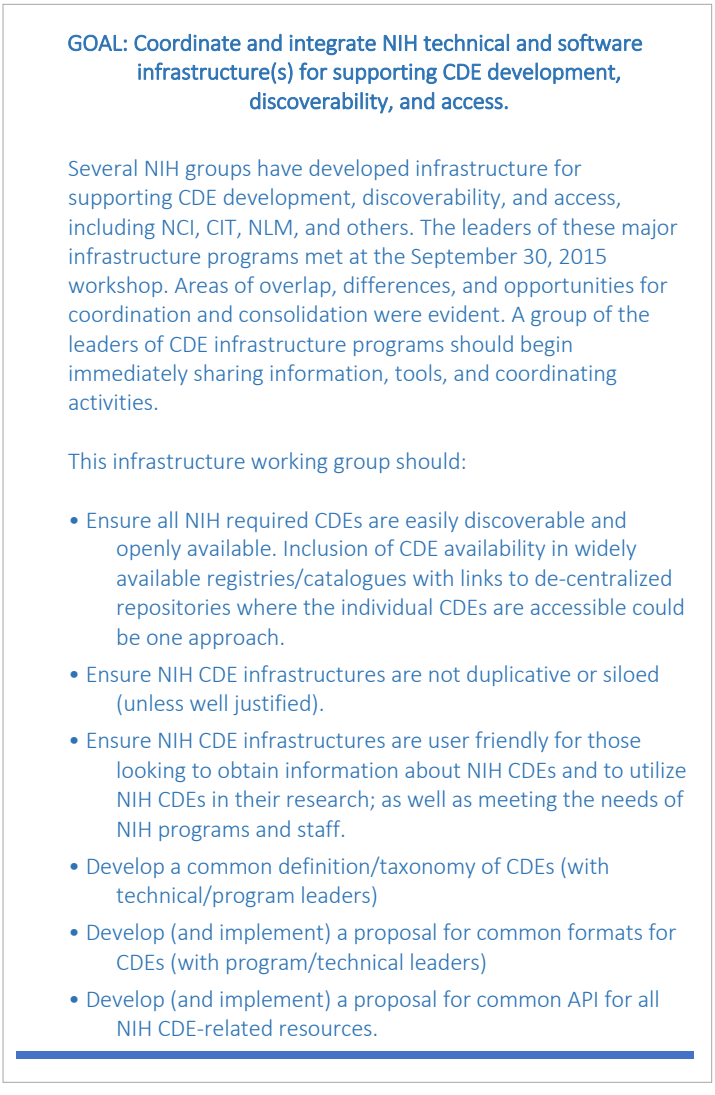


Figure 2: GOAL developed by the NIH Common Data Elements Workshop in 2015: Coordinate and integrate NIH technical and software infrastructure(s) for supporting CDE development, discoverability, and access.

The Working Group envisions the MRD as a "next generation" of the existing NIH CDE Repository currently developed and hosted by the NLM. The MRD could be created by developing additional features on top of the existing CDE Repository software platform, or by starting with and customizing a vendor-developed metadata registry software product.

The intent of this project is to have an NIH-wide reference library for key common data elements and forms, supported by tooling that can work interoperably with the diverse data repositories and metadata repositories also funded by NIH (by specific ICOs). This is intended to address the goal developed by the NIH Common Data Elements Workshop in 2015 (see Figure 1).

A fundamental goal of the NIH is to generate good and reusable data and to make data sharing meaningful. The MRD would help harmonize data elements and forms to that end. The Working Group recommends engaging subject matter experts to review and, where appropriate, help to standardize and harmonize NIH CDE content. For this effort to be successful, it is critical to establish a rigorous and well-defined governance process with clear scope and authority to set criteria for reviewing CDE content submitted by ICOs and promote the use of NIH-wide preferred CDEs. While respecting and enabling ICO-level subject matter expertise and focus, the MRD would also enable the NIH community to leverage shared technical infrastructure and other complementary shared resources.

Overall, CDE infrastructure supported through the MRD should have capabilities to support metrics collection to assess the quality of content. Adopting an appropriate standard for metadata registries is recommended, which can serve as an implementation guide for NIH ICOs during CDE infrastructure development work. To make CDE metadata usable and interoperable the MRD should incorporate standardized file format and APIs.

Software development should adopt and closely adhere to best practices to ensure usability of the MRD by both human and machine users. User roles might include Investigator, Grantee, Curator, Operations Team, and Program Official. Definitions of these roles are provided in Appendix A.

Importantly, adoption of CDEs requires more than just a technical software infrastructure solution. To complement the technical capacity of the MRD, the Working Group also emphasizes the need for ongoing work, led by the NIH CDE Task Force, to:

- Coordinate and develop educational and training resources about how to use CDEs in general, and how to use the NIH MRD, customized as appropriate for numerous stakeholders including investigators, grantees, program/project/scientific review officers, and CDE curators;
- Develop incentives for research initiatives to adopt CDEs (e.g. more widespread use of FOA language that strongly encourages or requires use of CDEs); and
- Determine key metrics and methods for tracking development, adoption, and use of the MRD and of CDEs through the workflow of research: at the study design level, during data collection, publishing, and for secondary use in data sharing platforms.

Functional Requirements (Core Services)

The Working Group identified and prioritized critical tasks and services as functional requirements of the NIH MRD for phases 1 and 2. The following five groups of functional requirements critical for phase 1 of the NIH MRD are categorized for convenience; however, some services relate to multiple categories.

Discovery, Search (Basic & Advanced), Retrieval, and Download

- From an authoritative portal, users can search/discover and retrieve NIH- and ICO-endorsed CDEs and electronic and paper case report forms comprised of CDEs.
- Search by both the individual data elements and forms themselves and by the studies/datasets with which they are associated, that is, by:
    - string/text
    - identifiers and versions
    - item's administrative attributes - e.g. permissible values
    - using wildcards
    - "related" search - "see also" to retrieve similar or related items
    - domain, condition/disease, and health topic (e.g. symptom science)
    - item's administrative attributes
    - ability to associate studies/datasets to form structures and/or CDEs, then search by those characteristics
- Display the prescribed form structure(s) and corresponding data elements as endorsed by the NIH
- Standard download of essential metadata attributes in commonly used representation formats including Excel, CSV, ML, and JSON
- Provide a means for grantees or investigators to extract information about the CDEs or forms being used to link to or upload the information into their own community web pages/site alongside information explaining how to appropriately/correctly use the metadata
- Support reproducibility of results: extract information about the CDEs or forms suitable for use in documenting a study to help investigators reproduce research results. Formats (e.g. Excel, csv, REDCap, SDC, ODM, FHIR, BRICS)
- Metadata driven user artifacts: extract or download CDEor forms information in formats suitable for use in data collection such as XSD, XML Document format, JSON files, Excel, BRICS
- Provide APIs and open interfaces in standard formats: provide access to NIH CDE and forms metadata through open and standardized interfaces (e.g. ISO or FHIR Data Element)
- Ability to find all forms that use a specific CDE
- Ability to define and save a customized download, selecting the specific metadata attributes and ordering that can be saved and reused
- Data platform CDE list - Generate a CSV file with all Data Elements (DEs) with 2+ studies using it. (to be able to monitor usage of different CDEs across NIH). This usage will also allow to spot the most significant and "winning" CDEs, e.g. those most used by principal investigators (PIs)
- Data platform study list - Generate an overview of studies included on the platform and whether they use any CDEs and how many they use. Distinguish studies in planning, ongoing, and completed stages
- A designation to characterize the study or initiative that used the CDE indicating the time period at which mapping to CDEs happened (before study, vs subject to post-hoc retrospective mapping), if applicable.

- Ability to link CDEs used in study or initiative to related persistent identifiers e.g. clinicaltrials.gov, data set in data repository, publications (PMID, PMCID, DOI), grant ID, FOA, etc.

Metadata Driven Software Development

- Export CDEs and/or forms and metadata to customize data collection tools to be used in different electronic data capture systems (e.g., CDISC, ODM, or FHIR platform agnostic format or platform specific formats such as formats of Medidata Rave, Theradex, Velos, Oracle Clinical, REDCap , and BRICS.
- Provide users with the ability to see related content, without tacit knowledge of the underlying ISO-11179 metamodel or specific information model or content in the repository
- Provide strong search algorithms and organizing structures to help more advanced users (especially curators) to identify collections of reusable content along with correct usage information
- Search for specific kinds of ISO/IEC 11179-3 registry components
- Metadata transport services: Bulk import/publish/export information about selected data element(s) in the NIH Dictionary
- Generate HTML forms
- Generate UML class diagrams

Vocabulary & Terminology Services

- Vocabulary services: create new CDEs linked to standard vocabulary and value sets to ensure that the MRD will be integrated with existing ontologies/terminologies
- Retrieve semantic code system cross-links (ability to integrate with UMLS to provide mappings to semantic code systems)
- Support the use of external and internal value sets

Registration, Submission, Governance, Audit, Curation, Harmonization, and Maintenance

- Submit and register Data Elements, Data Element Concepts, Value Domains, and Package Extensions
- Provide intuitive user interfaces, and use plain language
- Versioning and administrative lifecycle/workflow support (e.g. Draft, New, Retired/Deprecated)
- Capture administrative information for each entry in the repository such as notes, date created/ date updated, by whom
    - For a given CDE or FS – collect/record all history of audit information (e.g. date, notes, user who changed, etc...) to be able to identify when the CDE/FS was last updated or out of date (i.e. deprecated).
    - Generate audit reports.
    - Provide users of metadata with information about important changes.
    - Provide before and after values, as well as identifying related items impacted by any change.
- Clearly distinguish between NIH standards (NIH-endorsed) from ICO CDEs (ICO-endorsed). Ability to identify data element as a unique data element (UDE) or common data element (CDE)

- Registration lifecycle/workflow support (e.g., changes from “Recorded” to “Preferred Standard” status are tracked and trigger alerts).
- Ability to associate studies/datasets to form structures and/or CDEs
- Support user-defined validation/business rules for elevating the administrative or registration status of an item
- Create new content
    - Construct (or import to load) CDEs and forms from a standard/core set of metadata attributes
    - Create new content based on existing content
    - Dynamic extension: Create new CDEs aligned with or extending one or more national information models (e.g. BRIDG, PCOR). This requires the ability to register information models.
- Edit existing content
- Support bulk maintenance functions, such as adding additional names or classifications to existing content
- Model fragments to indicate relationships between data elements (e.g., blood pressure is composed of systolic blood pressure and diastolic blood pressure)
- Version control and tracking
- Harmonization Services: Provide services to help search, identify, compare, and harmonize similar CDEs. Enable harmonization and avoid duplication of content by providing alerts and support to identify and compare exact duplicates or near matches
    - Ability to identify and select similar items and view features and attributes of the items side-by-side.
    - Search/Identify related CDEs (age, age of onset of symptoms, age at diagnosis) and can compare details about context (self-reported, instructions for use)
    - Compare/search the level of detail of permissible values to determine appropriate level of depth. (Native American vs list of all 566 tribes recognized by US government)
- Support application of user-defined naming conventions
- For a given case report form (CRF) – obtain the endorsement priority/status and information about any licensing or copyright restrictions that indicate whether the form can be used without incurring any charges (i.e. is the form copyrighted-for-profit or copyrighted-for-preventing-changes)
- NIH should decide on a preferred information model to allow other infrastructures and systems to coexist with mapping between them (to achieve interoperability). This may be a daunting task. However, if we incorporate new data science technology, such as machine learning, the process may be streamlined and become efficient.
- Semantic web services: Generate semantic web content from a set of CDEs or eForms for use in publishing the content as a semantic web resources for linking to and exploring the meaning of data related to the CDE
- Provide the ability to leverage the semantics of CDEs to explore the meaning of the CDE, the meaning of directly related data, discovery of related data elements, data, and information using semantic web technologies, such as in a resource that leverages the associated terminology and uses Resource Description Framework (RDF) and/or other semantic web technologies
- Ability to add attachments (e.g. PDF, jpeg, etc.) to CDE or form structures. For example, the entry could link to a repository such as JIRA although access to that repository may be controlled.

- Ability to link the information that was used to create CDEs or forms. That means, it could link to a repository, e.g. JIRA. The access to this repository can be controlled.

Reporting and Administration

- Provide the ability to integrate with other ICO data dictionaries and generate/retrieve information for publication in NIH data registry or catalog service(s).
- Add “column(s)” to provide additional information about CDE attributes (e.g. types, title, CDE naming convention for various CDE attributes) to CDE or Form Structures (ref ISO 11179)
- Create new attributes to extend the metadata available for specific kinds of existing content without the need for software development
- Provide a mechanism for extending the repository to record additional information such as the ability to record structured information about research studies so that the published data and conclusions can be reproduced (e.g., BioCompute Objects Metadata)
- Reproducibility and reporting: identify CDEs that have been used for specific projects/trials/disease or that correspond to specific data sets in the public domain or based on publication reference.

Cost Estimate for NIH Master Reference Dictionary (a library of common data elements)

The Working Group estimated costs for the NIH Master Reference Dictionary based on ICO experiences with CDE and data registry initiatives, including input from ongoing NCI contract exercises for modernization of the NCI Metadata Registry.

The **cost estimate (see Table 1) covers the following tasks and services**, summarized here and described in detail in the Functional Requirements section above:

- Base Period Year 1 (see Table 1)
- Years 2-4 (see Table 2)
- Core services for creating, maintaining, browsing and downloading content
- Bulk import services
- APIs for accessing content
- Ability to organize content into groups
- harmonization activities through workflow support and rules
  Harmonization workbench & Curation Services Coordinating Center
    - NIH will need to provision/staff a harmonization services group to use these tools to curate, map, harmonize, negotiate, and align CDEs across ICOs
    - This includes establishing/selecting a high-level conceptual model to use in harmonization activities, to differentiate CDEs that are needed for different ICOs, but also organize content semantically
    - Based on experience of MRD WG members the estimated need for Curation Services Coordinating Center would be 9 FTEs.

The Working Group also identified several key considerations and needs **not represented** in this cost estimate:

- Training programs/materials offered in-person, online or through tutorials geared toward a variety of users.
- Ongoing outreach and customer service support. For example, a survey or documented user experience to improve the MRD over time.
- Additional tooling developed in years 2-3 to support operationalizing across ICOs
- Ongoing forum for representatives from ICOs who have CDE related and data registry systems, to share and converge ideas for improving tooling and approaches for interoperability and CDE support
- Some tasks will be manual initially but will need to incorporate computer-assisted or automated techniques over time
- As with many vendor products, the base year cost is significantly higher than the ongoing maintenance of the code base and support fee. The MRD is estimated at XXX at maintenance level plus the costs for hosting (whether by vendor, in cloud, or by NIH on premises)

Table 1: INDEPENDENT GOVERNMENT COST ESTIMATE (IGCE) for Base Period (Year 1)

NIH Master Reference Dictionary (a library of common data elements)

The Cost Estimate has been removed from this copy of the document which is posted on the NCI WIKI

**NOTE: Software development services estimates assume MRD is being developed from a partially-developed platform.*

***NOTE: Medical informatics expertise is a critical element of the needed subject matter expertise to ensure that the MRD work will be integrated with existing ontologies/terminologies.*

Table 2: INDEPENDENT GOVERNMENT COST ESTIMATE (IGCE) for Years 2-4

NIH Master Reference Dictionary (a library of common data elements)

The Cost Estimate has been removed from this copy of the document which is posted on the NCI WIKI

**NOTE: Software development services estimates assume MRD is being developed from a partially-developed platform.*
***NOTE: Medical informatics expertise is a critical element of the needed subject matter expertise to ensure that the MRD work will be integrated with existing ontologies/terminologies*

Appendix A: Definitions

Master Reference Dictionary & Data Elements Dictionary Infrastructure Working Group | Updated January 3, 2019

Terminology related to CDEs continues to evolve, and ICOs still use and interpret some of these terms differently. The Glossary Working Group of the NIH CDE Task Force standardized some of those terms during March 2017 to Jan 2018, with outputs available at https://github.com/lhncbc/CDE/tree/master/extra. Definitions standardized prior to March 2017 are also available at the NIH Common Data Element Resource Portal https://www.nlm.nih.gov/cde/glossary.html.

Term	Definition
Data Element* (DE)	Information that describes a piece of data to be collected in a study. The DE does not include the data themselves.
Common Data Element (CDE)	A data element that is common to multiple data sets across different studies. Commonality may be intentional or unintentional; this Portal places emphasis on the intentional use of CDEs to improve data quality and promote data sharing. Certain types of CDEs are sometimes described:
Form	An ordered set of data elements (or groups of data elements), instructions and rules that support the collection of specific information. Example: PHQ-9 Synonym: Form Structure
eForm	An electronic form used to collect data.
Embedded Dictionary	A Dictionary that is tightly bound to the data collected in a specific system; it would typically contain the data elements obtained from a reference dictionary
Reference Dictionary	A Dictionary that is independent of data but can be used for reference and/or data validation.
Data Repository	A system that stores data. Examples include a simple file system or a more complex system with or without access control that stores highly structured data, i.e., Data Lakes, Data Commons.
Informatics Systems	Sophisticated systems that have been engineered to provide tools to enable acquisition, validation, storage and processing of highly structured data. NIH examples include Federal Interagency Traumatic Brain Injury Research (FITBIR) & National Database for Autism Research (NDAR). External examples include electronic medical records and electronic health records systems.
Master Reference Dictionary (MRD)	A master repository or library that contains an authoritative and structured set of information describing the content, format, and structure of data. The base objects of the repository are CDEs and forms.

Stakeholder Role	Definition
Curator	One who manages data throughout its lifecycle, from creation and initial storage to the time when it is archived for posterity or becomes obsolete and is deleted.
Operations Team	Those who facilitates the application of Data Elements, Form Structures, or eForms in a system for data collection or submission

Grantee	A Principal Investigator (PI), Data Manager, Clinical Coordinator, or other agent who is required to apply Data Elements, Form Structures, or eForms for data collection or submission
Program Official	The individual responsible for all aspects of a Program(s) including budget, reporting and overall outcomes of project(s) to address an Institute's mission.

Appendix B: GSA Service Rates Used to Develop Cost Estimate

1 Program Manager, PM02 $115.29
2 Task Manager, PM01 $100.84
3 Data Architect (ER03) $122.85
4 Test Engineer (ER02) $122.85
5 Management Analyst (OM06) $99.87
6 On-Site Program Manager (OM05) $94.10
7 Sr. Program Manager (OM04) $72.37 $73.82 $75.30 $76.81 $78.35
8 Computer Systems Analyst V (CS05) $89.00 $90.78 $92.60 $94.45 $96.34
9 Sr. Systems Engineer (SY05) $113.04 $115.30 $117.61 $119.96 $122.36
10 Software Engineer Level II (SY02) $79.46 $81.05 $82.67 $84.32 $86.01
11 Application Systems Analyst/Programmer (SY01) $68.75 $63.51 $64.78 $66.08 $67.40 $68.75
12 Computer Programmer IV (CP04) $99.76 $101.76 $103.80 $105.88 $108.00
13 Computer Programmer III (CP03) $61.66 $62.89 $64.15 $65.43 $66.74
14 Security Engineer (IA03) $76.19 $77.71 $79.26 $80.85 $82.47
15 Network Admin / PC Technician (NT02) $49.09 $50.07 $51.07 $52.09 $53.13
16 PC Technician (NT01) $39.48 $40.27 $41.08 $41.90 $42.74
17 Subject Matter Expert II (SM02) $81.10 $82.72 $84.37 $86.06 $87.78
18 Help Desk Support (HD02) $40.21 $41.01 $41.83 $42.67 $43.52
19 Help Desk Coordinator (HD01) $33.57 $34.24 $34.92 $35.62 $36.33
20 Database Management Specialist II (DM02) $76.64 $78.17 $79.73 $81.32 $82.95
21 Database Specialist (DP05) $74.99 $76.49 $78.02 $79.58 $81.17
22 Technical Writer (TW02) $49.52 $50.51 $51.52 $52.55 $53.60
23 Document Specialist III (DC03) $26.99 $27.53 $28.08 $28.64 $29.21
24 Technical Administrative Support Specialist I (AS01) $30.47 $31.08 $31.70 $32.33 $32.98
25 Business Systems Analyst (BS02) $64.55 $65.84 $67.16 $68.50 $69.87
26 Functional Analyst (FA03) $81.07 $82.69 $84.34 $86.03 $87.75

Source: https://www.gsaadvantage.gov/ref_text/GS35F5457H/0OJBQO.36MJO2_GS-35F-5457H_GS35F5457HGPLOY.PDF

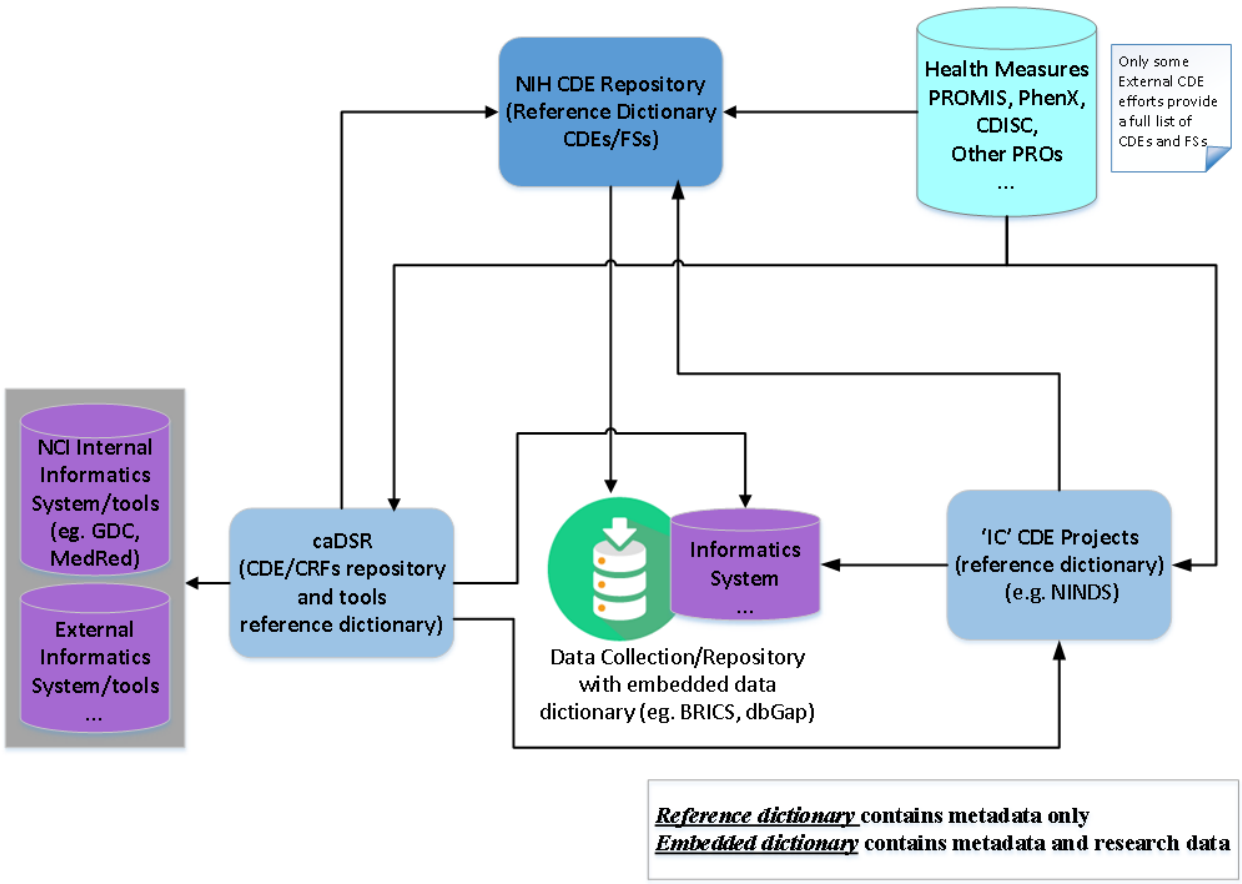
Appendix C: Figures from Infrastructure Working Group White Paper

Data Elements Dictionary Infrastructure Working Group | February 23, 2018

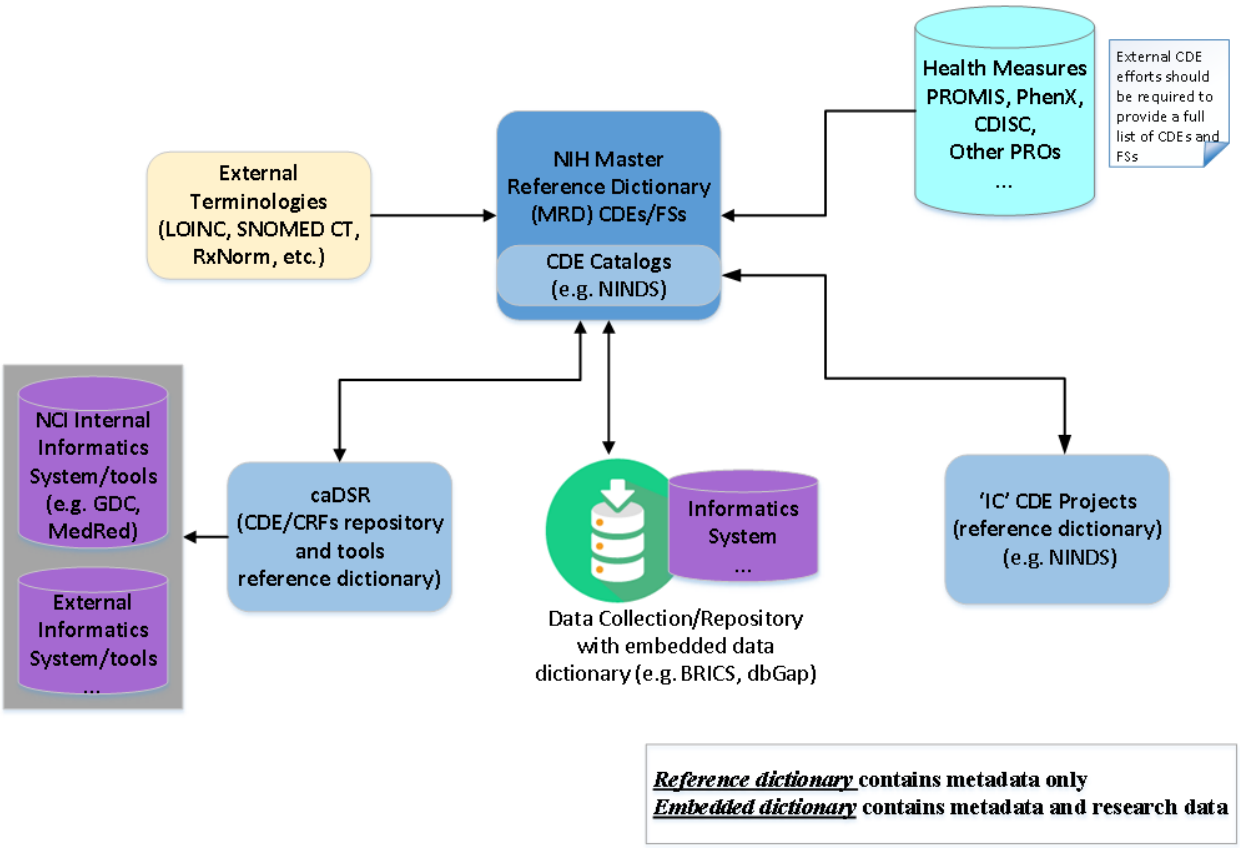
Matrix showing characteristics of CDE development and usage

		CDE Development	
		Centralized	Decentralized
CDE Usage	Centralized	• Data elements and data repositories are aligned • Streamlined submission process • CDEs are required in data repositories	• Investigators make decisions about metadata and data model • Data systems must manually map submitted elements to database dictionary (post-hoc harmonization) • Burden on data systems
	Decentralized	• CDEs are developed by IC • "Recommended use" • Investigators inconsistency (and/or inaccurately) use CDEs • No harmonization among variables	• Duplicate data elements are created; no linkage • Inconsistencies, inaccuracies, and factual errors are prone

Current Common Data Elements Environment



Proposed NIH Master Reference Dictionary to serve as the core component for proposed NIH CDE infrastructure (Phase 1)



Appendix D: Benefits of and Challenges to the Use of CDEs

Master Reference Dictionary Working Group | February 23, 2018

Benefits of and Challenges to the Use of Common Data Elements (CDEs)

USE OF COMMON DATA ELEMENTS IN GENERAL

BENEFITS	CHALLENGES
Use of CDEs (generally): • Facilitates interoperability, data sharing, and harmonization • Provides a mechanism for the collation and scale of smaller, isolated data sets • Reduces researcher burden in the creation of Forms and post-study analysis/harmonization • Facilitate secondary research & re-use • Reduces time harmonizing data • Reduces redundancy of effort • Potential to integrate research with Electronic Health Records	• Researcher Culture • Variation in the Format of data capture and analysis across disciplines • Resources and workforce to educate on CDEs and provide outreach to researchers • Lack of governance/enforcement mechanism • Need for consistent CDE validation requirements within and across disciplines • Misuse/Misunderstanding of CDEs • Lack of awareness • Researcher burden to discover, understand and implement

NIH USE OF COMMON DATA ELEMENTS AND THE NIH CDE REPOSITORY

BENEFITS	CHALLENGES
• The NIH CDE Repository could contribute to NIH's mission of supporting research and pioneering in data sharing • As the world's largest public funder of biomedical research, NIH can leverage this authority and brand to create standards that would be seen as reliable and facilitate re-use internationally • NIH could enforce CDE adoption throughout the IC's by requiring CDE enforcement in FOAs • A successful NIH CDE Repository may facilitate more integration and collaboration across research areas, both clinical and basic. This would be consistent with recommendations in the NIH Strategic Plan for Data Science about CDEs. • NIH-wide CDE initiatives, (including the NIH CDE Repository) could foster more collaboration between NIH Institutes	• Researcher Culture • NIH Culture – Differences between Institutes in definitions, metrics, and validation processes • Development and promotion costs to enact policy change and enforcement/incentives across Institutes • Promotion, education and outreach costs required to reach and educate userbase • Cost to incentivize education and usage of CDEs across ICOs • Differences in need and culture between basic research and clinical research communities • Variation in CDE validation standards across Institutes • Variation in CDE definitions across Institutes • Variation in formatting, metadata requirements, and intended purposes for CDEs across Institutes

Appendix E: Executive Summary from 2015 NIH CDE Workshop

September 30, 2015

1. Establish NIH wide CDE-related governance.

NIH should ensure there is NIH wide governance and oversight of CDE-related policy and coordination in the context of broad data sharing.

- It is recommended that this group report to the Scientific Data Council, who is responsible for data sharing oversight and policy.
- The NIH CDE Governance Group should address issues of NIH wide CDE policy and coordination among NIH ICs, with other federal agencies, the academic research community, and the private sector.
- The NIH CDE Governance Group should include representation from all NIH ICs to ensure NIH wide CDE coordination and policies enhance the work of all ICs.
- The NIH policies and practices should build on and incorporate the work and expertise of the existing BMIC CDE working group, which includes NIH staff with deep experience in multiple CDE programs.

2. Ensure NIH wide forum for sharing and harmonization across NIH CDE programs.

The Trans-NIH Biomedical Informatics Committee (BMIC) established a subcommittee of technical and program leaders of NIH CDE programs to share and document their experience in developing, managing, and providing CDEs for their research communities. This group has enabled information sharing across broad and diverse sets of CDE programs leading to an appreciation and respect for the commonalities and differences of various CDEs. This forum should continue, and these technical and program leaders should provide input and recommendations to the NIH CDE Governance Group concerning NIH wide policies.

This program/technical forum should:

- Develop shared common practices for CDE development, adoption, harmonization, modification, versioning, and use.
- Ensure appropriate flexibility such that NIH CDE programs can enable all types of research and address the needs of particular IC use cases and research. communities; however, encourage use of shared CDEs as appropriate.
- Develop a common definition/taxonomy of CDEs (with infrastructure leaders)
- Develop (and implement) a proposal for common formats for CDEs (with infrastructure leaders)

3. Coordinate and integrate NIH technical and software infrastructure(s) for supporting CDE development, discoverability, and access.

Several NIH groups have developed infrastructure for supporting CDE development, discoverability, and access, including NCI, CIT, NLM, and others. The leaders of these major infrastructure programs met at the September 30 workshop. Areas of overlap, differences, and opportunities for coordination and consolidation were evident. A group of the leaders of CDE

infrastructure programs should begin immediately sharing information, tools, and coordinating activities.

This infrastructure working group should:

- Ensure all NIH required CDEs are easily discoverable and openly available. Inclusion of CDE availability in widely available registries/catalogues with links to de-centralized repositories where the individual CDEs are accessible could be one approach.
- Ensure NIH CDE infrastructures are not duplicative or siloed (unless well justified).
- Ensure NIH CDE infrastructures are user friendly for those looking to obtain information about NIH CDEs and to utilize NIH CDEs in their research; as well as meeting the needs of NIH programs and staff.
- Develop a common definition/taxonomy of CDEs (with technical/program leaders)
- Develop (and implement) a proposal for common formats for CDEs (with program/technical leaders)
- Develop (and implement) a proposal for common API for all NIH CDE-related resources.

4. Obtain evidence of the impact and value of CDEs.

Evidence for the value and impact of CDEs on scientific advancement (positive or negative) is difficult to find. To obtain such evidence, NIH should support research to gather relevant data on impact of NIH CDE use on scientific research projects.

5. Assess the interoperability of NIH CDEs.

The ease of exchange of information collected using various CDEs is unclear. This is a very relevant question when studies use different CDEs to collect data on the same characteristic or activity, e.g., smoking. With multiple CDEs in use for collecting data on smoking, it is unclear whether these data are comparable or not. Additionally, various CDEs collecting the same data may collect it in different units or formats. NIH should support collection of data on the comparability and interoperability of the NIH required CDEs.

Appendix F: Infrastructure and Tools Group "Birds of a Feather" Summaries from the 2015 NIH CDE Workshop

September 30, 2015

The Infrastructure and Tools group met on Sept. 30, 2015 as part of the NIH CDE Workshop co-hosted by OD/ NLM/NCI/NCATS/NIEHS. Co-chairs: Duc Nguyen – NLM, Matt McAuliffe - CIT, Denise Warzel – NCI, joined by Vojtech Huser – NLM, Christophe Ludet – NLM. We were charged with sharing and reviewing details of the current infrastructure and tools that NIH CDE owners, managers, or aggregators are using in the various ICs; defining the differences and similarities related to features; user and manager experience; and expansion, cost and maintenance issues. We were asked to categorize the optimal set of features, usability, and functionality for each group of key users of resources and tools. The intended participants were owners and managers of infrastructure or tools used to develop, manage, or disseminate CDEs. We were also provided with a set of questions we should attempt to answer, but were only able to address some of these questions. Because not all ICs were represented, additional input may be needed to reach an NIH-wide consensus.

Each of the infrastructure teams presented 2-4 slides overviewing their system. These are available on the CDE Workshop wiki.

- What are best/optimal approaches for NIH support for infrastructure and tools for the development, management, and dissemination of CDEs? How should NIH provide the resources, tools, and infrastructure to research communities that are or will use CDEs? Should this be a centralized resource or federated resource?

Both approaches are technically feasible, and we also considered existing cultural differences. We agreed that the best approach was a hybrid, with a central registry for registration and archival, and decentralized repositories in ICs where they are needed. This will ensure that Content intended to be discovered and widely shared is centrally registered, like in a card catalog, or hosted by a central repository if the owner wishes to do so, with decentralized development and hosting of content in the ICs where its desirable to do so due to local operational processes and procedures, for example where the content experts reside. Also, some IC's may want to use the infrastructure to describe all of their data, even if not meant to be widely shared. Any decentralized IC repositories should be interoperable with the NIH Registry, to ensure that the standards meant be shared are registered in the NIH Registry where they will be accessible through a single interface, allowing ICs to keep some content purely local

We felt NLM and some ICs might be able to offer expert curation services across ICs, to help owners get content into NIH format. By centrally registering the content to be shared, search and retrieval for end users is simplified. By allowing content to be held at the ICs at their discretion, their ability to control the distribution of IC-specific content is also supported. The internal representation of content by ICs would not be mandated to be the same across all ICs, but for public sharing through the NIH central registry, ICs must offer a consistent standardized NIH API interface.

Whoever the source of truth is for specific content should have control over where the official copy is made available for broad sharing, and attempt to ensure that no other copies in other ICs are being made available through public interfaces. We did not agree on the criteria that would be used to

determine when and what content should be centrally registered, but we did agree that any data that is to be shared, must register the CDEs, protocols, and measures in the NIH Registry.

We did not determine how best to completely avoid creation of new, potentially duplicate, local content across IC in the hybrid environment, but discussed the possibility of supporting a federated query through common API interfaces for this purpose.

Optimally, duplicate standard content (e.g. PhenX) should not be available at multiple places, and should be avoided. Content might reside in an IC while its being developed, but it should not be registered at the NIH level until read to be shared. Retrieval of content should be possible through the NLM Registry, even if hosted by the IC for retrieval. Recommended formats include JSON, XML, CSV, and we agreed PDF is not considered a machine-readable format. Owners of content should be allowed to decide whether to have the content hosted by another IC, host it themselves, or deposit it at NIH.

- How can NIH promote best practices and sharing of tools and information among multiple infrastructures? Are multiple infrastructures needed and how should they work together?

We agreed that tools and infrastructure can be shared by ICs if so designed, and that some ICs and NLM might offer curation services to those who don't have infrastructure. This is a current model for both NLM and NCI where expert curators help curate content from raw documentation into structured machine-readable documentation. The infrastructures can work together through standardized NIH interfaces.

We realized through discussion that there are at least two different kinds of systems related to "CDEs" or Content, metadata registries and data dictionaries. We categorized the systems at NLM and NCI as "metadata registries and repositories". They hold descriptions of searchable data fields, data collection instruments or measures, and CRFs for retrieval and use by the ICs. They are very similar and currently provide duplicate and overlapping capabilities. Both support search, view, compare, download, curation of CDEs and CDE based form designs, but have slightly different implementations of the same structures and attributes. They use the same labels for many attributes but with slightly different meaning, which can be confusing to end users. For content development, both NLM and NCI require users to have accounts, but the rules for content creation are different. We should strive to have the similar rules around those attributes that are named the same, as well as content versioning and naming conventions, and common mandatory attributes across NIH, so that content is consistent for end users.

The CIT BRICS system contains a data dictionary used within the system to support data validation and data collection. The data dictionary is not intended to be accessed by external users to retrieve "CDEs" for reuse, though it does provide artifacts used by those submitting data to the BRICS system, and it has curation capabilities that are used to create new "unique data elements" (UDEs) within BRICS that extend the imported external standards. The BRICS data dictionary is shared by all the systems supported by BRICS so that duplication of CDEs across projects is avoided.

The Phenx Toolkit, NINDS, caDSR Downloads Page, EDRN portal represent additional kinds of publicly available systems in which IC's have made their content available to users for viewing and

downloading in different formats, for example PhenX and caDSR offer a way to get content in REDCap CSV DD format, and NLM Repository allows export in Structured Data Capture JSON format.

Another kind of system is the NLM CDE Portal provides an index and link to existing IC web sites where individual CDE initiatives are listed. IC project teams may have their own web sites that describe their preferred standards and how to use them. This is similar to a central registry feature where content is registered for discovery, but hosted elsewhere for retrieval.

We agreed we need to define the schema/minimum set of attributes that should be exchanged when registering and sharing content, to ensure that ICs can search and retrieve content in a consistent manner irrespective of how or where the Content is stored internally. It was mentioned that preferably the external interfaces/API would be based on some ISO standards, but extensions may be needed in order to satisfy NIH use cases.

Custom return formats should be offered that are suitable for consumption by systems widely used across the community, such as REDCap (CSV), and NIH Clinical Center systems (C3D CSV), and BRICS. These data systems each have different restrictions on data values, variable names, use of attributes such as unit of measure, minimum and maximum length, section titles, instructions, and branching logic, and many do not support tracking the identifier of the CDE used in their system. Another format supported by EDCs (e.g., Open Clinica, Formedix, Medidata Rave) is CDISC ODM format. Agreement is needed on how NIH content is to be transformed into these common formats such as name mangling rules, field type transformations, and handling of unique identifiers.

Moving forward, we should seek to share reusable infrastructure, designs, and code rather than developing duplicate functionality where possible. Even if an IC wishes to develop and host their own content, the software that is needed can be designed and developed based on common requirements, and implement common interfaces. For example, once we agree on the common attributes of CDEs to support form design and development of measures and protocols, then by joining forces to define requirements, a single form design tool for those ICS that wish to make standard CRFs available based on CDEs could be developed that could be used to import CDEs from the NIH registry or any IC's repository of CDEs.

Best practices and sharing can be facilitated by the use of a common lexicon for discussing CDEs, forms (or instruments), protocols, and measures (currently referred to as "CDEs"). The meaning of some of these terms could be drawn from existing IT standards where they exist. Some terms may need to be constrained to use controlled terminology to represent their expected values such as "status" and "state", for example terms like "Qualified", "Standard", "Candidate" should mean the same thing across ICs.

Other terms that would benefit from standardizing: CDE, Variable, Form, Form Structure, Branching logic, Calculation Logic, Computer Assisted Testing, Survey Form, Repository, Registry, Data Dictionary, Global Library, Local Library, Portal, Unique Data Element (UDE), Value Set.

We also agreed there is a need for consistent subscription services to notify end users of details around important changes in order to support IC change management processes. Currently, NCI provides detailed reports for CDEs and CRFs formatted and delivered in machine and human readable format as an HTML report delivered via email to curators and as a file for processing by automated servers.

An ongoing infrastructure and tooling working group is recommended to ensure sharing of information and tools to the degree possible, all ICs should be invited. The host will rotate based on volunteers, to ensure active participation and balanced views of requirements and solutions.

- How do resources and tools interface with other CDE infrastructure outside of NIH? How does NIH CDE infrastructure incorporate other standards that are required, such as from ONC, CMS, FDA, or domain specific technical standards (e.g. radiology)?

Both NCI and NLM note that tools for curation of content intended to be shared broadly need to implement interfaces to consume of NIH terminology and standards, as well as relevant and preferred national and international standards. Often an IC may have content that was not developed using external standards. This content will need to be edited (possibly by NLM) before it can reach an NIH-preferred level.

NCI already provides this capability for use of external terminologies that are offered in machine readable format through APIs, and plans to expand it to reuse NLM and other new resources offered by NLM, CDC, and CDSIC. NLM has similar plans for the near future. We think that standards for NIH-wide sharing will need to be reviewed and finalized in a system with these capabilities to incorporate and harmonize with these external standards where appropriate. Some IC's may not have the resources to do this, so those with available infrastructure must provide these capabilities as services to other ICs to help transform IC–specific content into NIH-suitable formats.

Appendix G: Summary of Systems & Standards Presentations

Dictionary Infrastructure Working Group | February 23, 2018

23 May 2016: NCI caDSR Dictionary: Denise Warzel: The NCI Cancer Data Standards Repository (caDSR) is a reference dictionary that achieves these goals: A repository and tools to create and share well-vetted and community approved data element and caser report form (CRF) standards; Reduce time to set up a new protocol by reusing existing CDEs/CRFs; Reduce the cost to collect and analyze NCI clinical trial results; Improve consistency of data across NCI Clinical trails through the reuse of data elements; Metadata driven customization of clinical data collection systems

6 June 2016: BRICS Dictionary: Matthew McAuliffe: The Dictionary module is the cornerstone to BRICS data sharing. All BRICS data is tied to data elements, form structures and electronic forms, enforcing common vocabularies and data continuity within and between programs. This distinguishes the BRICS Dictionary from most workgroup Dictionary tools that are not tied directly to data submission, as changes to a data element in BRICS will have a direct and significant impact to submitted data and the ability to align data in queries.

18 July 2016: BRICS Platform: Matthew McAuliffe: The Biomedical Research Informatics Collection System (BRICS) offers research programs a secure platform and suite of tools to support data definition, data contribution, and data access throughout the research life cycle. The BRICS platform boasts current technologies and adoption of industry standard protocols for data transfer, as well as prominent features that include live data capture, semantic data query, biospecimen ordering, Terabyte file

8 August 2016: NIH CDE Repository (NLM Dictionary): Christophe Ludet: Demonstration included Forms and Data Elements from several systems, integrated into several systems, and topics surrounding metadata and concept codes

22 August 2016: BTRIS Data Warehouse: Jose Galvez: BTRIS is NIH CC 's clinical data warehouse, clinical research data is put into BTRIS for secondary research analysis across clinical data. RED is the "Research Entity Dictionary", it stores a concept for every data point managed within BTRIS

26 September 2016: Ptolemy V. Application: Bill Tulskie: Ptolemy is a tool for mapping data to standards and combining data from different sources. Data elements can also be mapped from preloaded NIH data elements. Ptolemy supports uploading data from csv, where it is analyzed by the tool in a process to identify potential data element matches. These potential matches can then be assigned to the source data. Following this process, the user may enter the data elements to be assigned to the data that were not automatically identified.

24 October 2016: IS0/IEC 11179 Standard: Denise Warzel: ISO/IEC 11179 is a 6 part metadata registry standard. Part 3 is the most complex, and describes the registry metamodel and basic attributes, which includes a metamodel for Concept Systems, Value Domains, Data Elements, and Data Element Concepts. Primary aims include providing a means by which to register a set of attributes for data elements that ensure the meaning of data is accurate, unambiguous, controllable (versionable), and verifiable, with focus on attributes for the creationg and management of these descriptions, then on the attributes required to use the data. It is designed for humans and machine processing, and supports ontological

reasoning of data from ontology definitions. ISO/IEC 11179-6 describes the procedures by which *metadata item*s can be assigned internationally unique identifiers and registered in a 11179 metadata registry that is maintained by one or more Registration Authorities. This standard includes a generic approach for the representation of any type of item in a metadata registry by its registration record that documents the common administration and identification, naming and definition details together with their metadata item-specific details.

Appendix H: High Quality, Reusable Data for Biomedical Research: Introduction to Clinical Common Data Elements (CDEs)

NIH CDE Task Force | February 20, 2018

For reasons widely acknowledged (Wilkinson et al., 2016), it is important that data and other digital objects representing the products and processes of modern biomedical science, are Findable, Accessible, Interoperable, and Reusable (FAIR). The use of standards, such as common data elements (CDEs), is key to the interoperability and reusability of data, allowing data to be more easily analyzed, shared, and combined with other data to derive knowledge and accelerate discovery.

CDEs facilitate the reuse of common variables and measures to provide consistency across multiple data sets, and encourage the connections between research and clinical encounters. CDEs make the data meaningful by structuring and defining commonly used, community shaped, recommended measures and assessment instruments. Through incentives made via funding announcements, awards, contracts, and other resources, the NIH supports the use and development of CDEs in clinical studies, patient registries, and human subjects research.

While there are varying levels of maturity and adoption of CDEs, there is general trans-NIH consensus about the importance of using them. To streamline communication and coordination across the NIH, the trans-NIH Clinical Common Data Elements Task Force (CDE TF) was formed. The CDE TF, now reporting to the Scientific Data Council (SDC), is uniquely positioned to assist with providing context about what CDEs are, why they are important, and lessons learned from the implementation of numerous CDE initiatives across NIH. For anyone considering recommending or using CDEs to collect biomedical data, this short document provides basic background. (While CDEs are discussed here as they pertain to data collected from human subjects, CDEs could also be developed for preclinical studies.)

What is a Common Data Element (CDE)?

A CDE is "a combination of a defined variable paired with a specified set of similarly coded responses to questions that are common to multiple data sets or used across different studies. CDEs are used in research where measurement, reproducibility, and comparison across studies is important. They can be structured as a single data element, or may be included in a collection of data such as a survey scale. Use of CDEs can facilitate data sharing and standardization to improve data quality and enable data integration from multiple studies" (Sheehan et al., 2016; also see Corwin et al., 2017).

When should CDEs be used?

Although CDEs might have been originally developed to address the needs of a specific research domain or clinical care application, many CDEs address universal concepts of interest to a wide variety of domains for a variety of data collection purposes, such as demographic characteristics of research participants. Identifying and reusing existing CDEs paves the way for smoothly finding, interpreting, and exchanging data.

CDEs can be used to promote data sharing, meta-analyses, and reuse. For example, to search across different data sets and identify a population with shared characteristics such as common diagnoses, symptoms, procedures, or interventions.

CDEs and other data standards can improve efficiency and quality of data for any information that is: collected frequently within or across domains, part of a validated assessment instrument, captured in an electronic health record (EHR), provided to comply with regulatory requirements, or used to conduct evaluation. For very frequently used data elements such as race, ethnicity, country of birth (ISO codes), and some patient-reported outcome measures (the PROMIS questions), there are widely accepted (i.e., common) existing CDEs that should be used in most instances. CDEs can be developed for information specific to only one or a few domains, such as particular neurological disorders or rare diseases, when stakeholders work together in a transparent and inclusive manner to build consensus around data parameters. However, even these "domain-specific" CDEs often transcend multiple domains. For example, CDEs used to reflect cognitive status in Parkinson's Disease might be suitable for studies of other conditions.

CDE Stakeholders
• Academic institutions • Clinicians • Data repository developers • Data repository specialists • Data scientists • Industry • Informaticists • Investigators • Librarians • Research Funders • Terminologists • Patients/Public • Policymakers • Statisticians

Who can CDEs help?

Health data comes from many different sources. Many different people produce this data, fund its collection, curate it, use it, and establish guidance around its collection or use. These include patients, funders, research participants, investigators, clinicians, industry, data repository specialists, librarians, informaticists, data scientists, policymakers, and the public.

To be able to share data across different sources, they need to speak the same language, or translate their language to a "*lingua franca*" – a common language all can understand. The CDEs are that common language.

How can CDEs help us get high quality, reusable data for biomedical research?

- Provide a common understanding of the data and make it FAIR. Use of CDEs ensures consistency across multiple data sources (e.g., studies, registries, EHRs) and allows for seamless exchange, combination, aggregation, harmonization, comparison and combined analysis of data among researchers and clinicians without the need for mapping. For example, if linked to the same codes to store the data "behind the scenes" in the database, there would be seamless interaction between a study that records patient sex as "male" or "female", and another study that enters the data as "M" or "F."
- Reduce time and cost needed to design a study, and develop data collection tools and repositories. CDEs provide a library of data points that can be used to create data collection forms. Valuable time can be spent determining *what* data needs to be collected instead of *how* the data should be collected. Some CDEs even have accompanying toolkits to make study design and data collection easier for investigators.
- Ensure regulatory compliance. Increasingly, regulatory agencies such as the Food and Drug Administration require compliance with specific standards. By using these standards as the foundation for data collection and storage, compliance is ensured from the beginning of a trial, and the reporting process becomes easier and does not require data re-entry.

- **Bridge clinical encounters and research.** With the growing availability of EHR data and the evolution of the Clinical Research Informatics (CRI) field, tremendous potential exists for leveraging EHR data to answer complex medical research questions. But to fulfill this potential, CDEs are needed to standardize data collection from research and clinical care both across and within disease domains. The benefits of such standardized terminologies are recognized through requirements for Medicare and Medicaid reimbursement and EHR Incentive Programs.
- **Accelerate advances in research.** CDEs are a key aspect of data-powered health. By using CDEs when first collecting biomedical data, it is easier to develop meaningful analyses and biomedical research projects. The data associated with CDEs can be more readily analyzed and reused to accelerate research to better understand the pathogenesis of diseases and mechanisms of health, and can be used to develop therapeutics to improve quality of life.

What are the challenges in developing and selecting CDEs?

Initial development of CDEs can be resource intensive. Significant effort is needed to reach agreement across the variety of stakeholders interested in a specific topic area to determine what questions to ask and how to ask them – for example, what are the permissible values and will they be presented as text or multiple choice? This may require reaching consensus about balancing the time and costs of collecting data with the need for that data, and these value determinations may vary over time or for different stakeholders. In addition, the learning curve for implementing CDEs can be high. What's more, some researchers may feel limited or burdened by requirements to use specific CDEs. One of the biggest challenges for implementation and continuous use of CDEs is governance. Governance includes formalizing ownership, maintaining versioning history, providing different types of role-based access (editing to exporting), to assure that the encoding is accurate and evolves in sync with changes in biomedical research and clinical practice.

CDE Advantages
• Accelerates advances in research. • Reduces the time and cost needed to develop data collection tools and repositories • Promotes standardized, consistent, and universal data collection • Improves data quality • Facilitates data sharing • Improves opportunities for meta-analysis and comparison of results from different studies • Enhances community engagement and the importance and understanding of data science • Promotes public-private partnerships

Overcoming challenges.

While CDEs may not currently exist to meet all data needs, use of existing CDEs decreases the preliminary workload and increases the potential quality of the data. Where there are gaps in CDE coverage, a CDE champion can help to organize stakeholders and facilitate productive collaboration. Involving industry, patients, and the public as partners in various efforts will enhance community engagement and emphasize the importance and understanding of data science, as well as promote public-private partnerships. Support from stakeholders in developing and adopting CDEs can help to reduce the time and cost needed to develop data collection tools and repositories. Investing this effort to create a shared resource may be costly at the outset, but will prevent the need for each initiative to independently undertake the same process.

Electronic data capture tools, which use computerized systems to collect data in electronic rather than paper form, can reduce the overall time required to capture, store, access, exchange, and compare data.

Such systems rely heavily on the computable information associated with the CDEs to generate a graphic user interface for data entry, run algorithms based on entered data such as computing a score for a survey instrument, apply validation rules to check data quality, and enable reporting functionality to analyze and export collected data.

The challenges involved with governance can be addressed by using appropriate tooling to provide a structured framework and functionality for versioning, using standardized terminologies, and providing appropriate attributes for defining ownership, roles, export options, and more. Use of existing tooling, such as the NIH CDE Repository, and existing CDEs, where possible, can decrease the time and cost burdens to investigators so their own resources can be used on other items.

Building upon best practices

Based on the experiences of many data standardization projects conducted or funded by NIH, the CDE TF compiled a list of lessons learned and suggestions for promising practices to help NIH initiatives develop and adopt CDEs.

References

Corwin, E. J., Moore, S. M., Plotsky, A., Heitkemper, M. M., Dorsey, S. G., Waldrop-Valverde, D., Bailey, D. E., Docherty, S. L., Whitney, J. D., Musil, C. M., Dougherty, C. M., McCloskey, D. J., Austin, J. K. and Grady, P. A. (2017), Feasibility of Combining Common Data Elements Across Studies to Test a Hypothesis. Journal of Nursing Scholarship, 49: 249–258. doi:10.1111/jnu.12287

Sheehan, J., Hirschfeld, S., Foster, E., Ghitza, U., Goetz, K., Karpinski, J., Lang, L., Moser, R.P., Odenkirchen, J., Reeves, D., Rubinstein, Y., Werner, E., Huerta, M. (2016) Improving the value of clinical research through the use of common data elements. Clinical Trials, 13(6), 1–6.

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016).

Appendix I: Common Data Elements: Promising Practices and Lessons Learned from NIH Experience

NIH CDE Task Force | February 13, 2017

Across NIH, a number of ICs are actively engaged in identifying and encouraging the use of common data elements (CDEs) by funded investigators. A CDE is a fixed representation of a variable to be collected and analyzed in projects within a particular analytic or clinical domain. A CDE consists of both a precisely defined question and a specified format or set of permissible values for responses (e.g. answer list). Ideally, a CDE is defined unambiguously in both human and machine-computable terms. Sets of CDEs consisting of individual question/answer pairs can be combined into more complex questionnaires, survey instruments, and case report forms for use in collecting data for clinical care, research, patient registries, and surveillance studies.

CDEs offer significant benefit to biomedical research. More widespread use of CDEs can accelerate the start-up of new research projects by providing investigators with a set of established guidance for collecting data about certain variables of interest. CDEs can contribute to improvements in the quality of data collection by promoting the use of validated data collection instruments that contain them. CDEs can advance science by facilitating data sharing, comparison of results across research studies/diseases, interoperability to enable the aggregation and analyses of data from multiple studies/registries to provide new insight and/or greater statistical power. Common, precise and unambiguous definitions of variables are a necessary first step to aggregating data that is collected in different studies or registries. CDEs that are specified using standardized vocabularies, code sets and terminologies can potentially ease the burden of data collection for projects that derive some of their data from clinical encounters, assisted by a certified electronic health record.

CDEs are generally identified and developed by organizations and groups of subject matter experts in particular fields (e.g., neuroscientists, hematologists) that determine that the specification of the CDE is fit for intended purposes. A CDE created for one analytic or clinical domain may nevertheless be useful in other contexts, and CDEs identified in different domains may be identical or closely enough related to merit possible harmonization. A number of NIH ICs have invested in identifying CDEs for use in a variety of domains, from substance abuse to neurological disease, to rare-disease patient registries and to genome-wide association studies. Other ICs have expressed interest in identifying or developing CDEs for use among their supported research communities. NIH CDEs are envisioned for use in and across multiple settings, such as clinical studies, data repositories, patient registries, and basic, applied, and epidemiologic research.

This document presents a set of lessons learned from NIH experience to-date in identifying and developing CDEs for use in NIH-supported clinical research, patient registries, and other human subject research. It is not intended as a comprehensive, step-by-step guide to identifying/developing CDEs, but as a select set of good practices that can be used in future NIH

CDE initiatives, or in other national or international projects. The lessons do not provide guidance for determining *whether* to identify CDEs for a particular domain, but focus instead on *how* to implement initiatives to identify and promote the use of CDEs. The document is organized around twelve key lessons related to project initiation and formulation, data element selection, and efforts to promote uptake and use.

The document was originally prepared by the CDE Working Group of the trans-NIH BioMedical Informatics Coordinating (BMIC) Committee. The CDE WG was the precursor to the CDE Task Force of the NIH Scientific Data Council. This document will be updated periodically to reflect continued NIH experience with CDEs.

Lesson 1: Don't reinvent the wheel – Determine what information you wish to gather, and whether data elements of interest already exist.

Before launching an initiative to identify/develop CDEs for a particular purpose, it is important to review data elements that are currently in use to avoid unnecessary duplication and the proliferation of competing data collection approaches. Data elements have already been developed for a wide range of diseases, conditions, symptoms, and purposes; some have been designated as standards and/or achieved widespread use. A CDE from an existing initiative may be of value to a broader set of topics; the most "common" CDE would be useful across broadest set of studies/registries/uses. We recommend starting out with a small pilot project to learn the process, then expanding.

The CDE Resources Portal provides a useful starting point. Data elements listed under "NIH CDE Initiatives" are those that have already been identified for use in certain (specified) types of NIH-supported research and patient registries. If data elements relevant to a particular field or study of interested are not listed in the Initiatives, candidate data elements may be found in the list of "NIH CDE Tools and Resources," which identifies larger sets of data elements and case report forms that are frequently used in research settings, along with tools for searching them.

Allow a mechanism for grouping CDEs into sets (consider using Form Sections in your forms whenever appropriate). The NIH CDE Repository provides this capability.

Other resources you may wish to consult when starting a CDE initiative are ClinicalTrials.gov, - the Database of Genotypes and Phenotypes (dbGaP), and NCI CDE Browser. ClinicalTrials.gov contains summary results of thousands of clinical trials. It can illustrate how researchers measure and define relevant outcomes in different disease areas and assist in determining which measures and which ways of collecting them are most common. DbGaP contains results of genome-wide association studies and provides access to the protocols used to define phenotypic measures collected in the study. Some studies in dbGaP make use of NIH-supported CDEs, including PhenX. The NCI CDE Browser contains all the CDEs used in all of NCI's clinical trials since 2002, as well as CDEs that have been vetted and standardized across all of the NCI Divisions, Offices and Centers who use CDEs. The browser also contains a link to "Downloads" that include more than 60 standardized NCI Case Report Forms (CRFs) containing NCI CDEs.

It is also helpful to engage outside experts representing disparate disciplines, as well as those in other appropriate federal agencies or ICs early in the process, particularly for those initiatives containing multiple disease areas or domains.

Lesson 2: Don't work alone - coordinate with other ongoing efforts

Many CDE initiatives are already under way. Organizations and institutions internal and external to NIH have an interest in identifying/developing CDEs for similar diseases or conditions, or for multiple populations (e.g. social determinants of health). Joining forces with other likely collaborators is a good step toward greater standardization, greater commonalty, and greater adoption of a set of CDEs. The CDE Task Force aims to identify CDE and other data standards initiatives across NIH and beyond. The CDE Task Force established an internal CDE Wiki that can be used as a platform for announcing interest in a particular domain and seeking collaborators within the NIH.

Outside NIH, many organizations are involved in identifying CDEs. The FDA, for example, is working with the Clinical Data Interchange Standards Consortium (CDISC) and other organizations to develop data standards for FDA-regulated clinical research related to some 60 designated therapeutic areas. The CDE Task Force is coordinating NIH input into that process, with the aim of developing data standards that will be useful across the spectrum from academic to industrial research. Information on the FDA initiative and NIH contributions to it can be found on the CDE Wiki.

Even within a single CDE initiative, coordination mechanisms are needed to promote consistency across subgroups that are established to identify CDEs for related subdomains or disease areas. An overarching coordination function can help reduce redundancy and "near misses" across the initiative and foster convergence on more limited (single?) ways of collecting certain types of information of interest to multiple subgroups.

Lesson 3: Engage the full range of needed expertise - domain experts, informatics experts, community experts, and other relevant experts as the content changes.

Effective development/identification of CDEs requires a range of expertise. Expertise in the relevant research domains (e.g., cardiovascular disease or neurological disease) is necessary to identify measures of primary interest and assess their validity and viability in practical settings. Such expertise can come from scientific researchers, clinicians, nurses, and other health professionals, each of whom brings a unique perspective to the process. Expertise in informatics is necessary to develop or select data elements that are consistent with existing standards, including those used in clinical care settings and electronic health records (EHRs), to define data elements in specific measurable terms, and to express data elements in ways that are syntactically and semantically interoperable. Representatives of the patient community (e.g., patient advocates) can also bring valuable expertise and perspective, e.g., in identifying the

measures of greatest interest to patients and in considering practical issues of data collection and administration. The most effective CDE initiatives include from the outset the full range of expertise needed for their project. A core group (Steering Committee) should be established with central expertise such as informatics specialist, and core leadership; and then as the subject matter changes, ask for expert(s) to join until CDEs have been established.

Lesson 4: Use data elements that are scientifically validated, whenever possible.

CDEs should be brief and take into account the burden of associated data collection on the patient/subject and on the clinician/researchers, as well as ease of use – especially when choosing between similar questions or instruments. When identifying data elements for inclusion in a CDE initiative, it is preferable to select data elements that have been tested to establish their validity, reliability, sensitivity, and specificity to the disease or condition of interest. For establishing new CDEs, efforts should be made to validate data elements across the populations of interest, taking into consideration characteristics such as race/ethnicity, socioeconomic status, or geographic areas that may be involved in a study.

In many cases, fully validated data elements will not be available for all variables of interest. In such cases, it is important to have a clearly defined process for selecting data elements (e.g. use of expert consensus) and to document this process for others to see. Some CDE initiatives classify data elements into multiple categories, such as required and experimental or core, supplemental, and emerging, to signal differing degrees of validation or consensus.

Efforts should be made to provide potential users with information and guidance about the validity and provenance of designated CDEs. The validity of a data element can be highly dependent on factors such as the context for administration (e.g., in a primary care setting) and on the patient population (e.g., adult vs. pediatric patients). The CDE Portal includes a brief summary of the validation process (if any) and provenance of data elements, as well as a link to a foundational publication that may include additional information about validation. Some initiatives provide more complete information about validation and provenance on their websites. Making sure such information is easy to find is essential to proper use of the CDEs.

Lesson 5: Use data elements that conform to existing data standards and are freely available, whenever possible.

It is preferable to use data elements that conform to existing data standards (e.g., terminology standards) and regulatory requirements in their domain of use. Data elements need to be consistent with current regulations and standards specific to the region or agency they serve. CDEs that express questions and answers using clinical terminologies that have been identified for use in certified electronic health records under the U.S. program for meaningful use, for example, will potentially improve the ability to compare data collected across different

electronic health records systems. Many of these data standards are also used internationally (e.g. LOINC and SNOMED CT), and can facilitate comparability with data collected in multiple countries. LOINC contains a number of survey instruments and CDEs used in NIH CDE Initiatives, including PROMIS, Neuro-QOL, and PhenX, in addition to lab test and clinical measures, and the LOINC website allows searching across these available CDEs. SNOMED has been mapped to ICD. For research studies including clinical trials, CDISC standards, which are based on NCI Thesaurus terminology, have been promoted by some U.S. and international regulatory agencies and other organizations (e.g. for reporting clinical trials to the FDA). The NCI Thesaurus is also used in other research contexts -- both directly and through CDEs.

Preference should be given to data elements that are freely available for use, with minimal licensing restrictions, to facilitate use of identified CDEs. Data elements that are only available for a fee or with restrictive terms of use can impede uptake and add to the cost of doing research. In some cases, ICs have paid licensing fees to permit the use of copyrighted instruments by funded investigators; in other cases, ICs may direct users to the copyright owner to make arrangements for individual use of the designated instrument. In the latter case, the CDE website might only be a link to the instrument or paraphrase a question rather than provide the full instrument.

A few important specifics and characteristics of CDEs:

- Define a clear data type for each CDE and use a rich set of data types (e.g., boolean, integer, positiveInteger, decimal, quantity, time, date, datetime, CodeableConcept) https://www.hl7.org/fhir/datatypes.html#primitive
- Annotate each of the individual members of a value set with a common terminology (e.g., SNOMED CT, NCI Thesaurus, MedDRA, LOINC, or other relevant terminology. In other words, annotate not only the question text, but also each of the possible answers (e.g., if it is a drop down list question).
- Use a consistent strategy for capturing why a CDE value is missing, known as "flavors of NULL" (e.g. 'Unknown' vs. 'asked but unknown' vs. 'not asked'). The HL7 FHIR Value Set for this is available at: https://www.hl7.org/fhir/v3/NullFlavor/index.html

Lesson 6: Plan ahead to update CDEs over time and keep track of all versions.

CDEs need to be reviewed periodically to assure they remain valid and reflect the current status of the disease or condition. Updates are necessary to reflect new science, experience with existing data elements, changes in practice, expansion of the set of users of a CDE, or changes in other data standards that are embedded in a data element. All Versions of the CDEs should be maintained so that users can retrieve accurate information about the version of a CDE that was used for a particular dataset. Early in the process of planning a CDE initiative, consideration should be given regarding the frequency required to review, update, and release new versions of

identified data elements. Consideration will also need to be given to procedures for allowing cross-walks between newer and older versions of data elements. Revisions and updates can also have spill-over effect on related CDE initiatives that make use of the same or similar data elements and on the systems used by researchers to collect process research data. Update efforts should include participation from such other affected groups and institutions. All of these considerations should be part of the charge to the core group or steering committee, and also added as a policy to governance documents.

Lesson 7: Gain support of the user community.

CDEs are only as useful as the intended user community perceives them to be. Securing the support of the user community and defining their role in the process of development is paramount to the successful application of CDEs. Enlist experts in the field and patient advocates for their input during CDE development to facilitate this support. Given the time, effort, and cost associated with CDE efforts, it is also important to get the buy-in of institutional (IC) leadership and ensure they understand the potential benefits of the initiative. Including these stakeholders on the CDE development team is a good approach. Social medial can play a role in this context. The GEM (Grid-Enabled Measures) database, for example, uses social media to allow members of different user communities to rate measures used in particular domains and develop consensus around them. Social media tools can also increase exposure, create communities, and advance the meaning of CDEs.

Lesson 8: Develop a communications strategy

Consideration must be given to determining how the relevant research communities will learn about and find the CDEs developed/identified for a particular field of research. Journal articles and press releases can be effective for announcing the availability of CDEs and explaining their significance and potential uses. Engage the community by poster presentations or podium presentations at large scientific meetings. Officially issued funding announcements (FOAs) and Requests for Information (RFI) can also be a platform to increase awareness, as well as encourage or even require the use of a prescribed set of CDEs as part of the data management plan. Social medial venues can also be effective in disseminating information about CDE initiatives to communities of interest. Collaboration with professional organizations can also help with communications and with gaining user support. Additional steps are often needed, however, to provide a more stable, permanent, and visible source of information about the CDEs. Listing CDEs in the NIH CDE Resources Portal is one way to enhance awareness of the CDE initiative and access to the data elements. Because the Portal provides only a summary level view of the initiative, development of a website specific to the CDE initiative can provide a complementary source of information about the data elements and a platform for further communication and access to updated CDEs. Most of the CDE initiatives listed in the CDE Portal have Websites explaining the initiative and providing direct access to the data elements. Data elements themselves should be deposited in repositories designed to provide long-term access to relevant communities of users.

Lesson 9: Develop policies to promote the use of CDEs

The availability of CDEs will not, by itself, motivate investigators to use them. Specific policy language is often needed to encourage, expect, or require NIH-funded investigators to make use of CDEs relevant to their research. Many ICs have developed language for inclusion in Requests for Applications, Program Announcements, and Terms of Award to compel the use of designated CDEs in particular types of research. Examples of such language are provided in Appendix A. In addition, the CDE Task Force developed language that is now contained in the NIH Program Template now (in section IV.6) and can be inserted into Funding Opportunity Announcements to encourage the use of relevant CDEs in NIH-funded research, where appropriate. That language is also contained in Appendix A and has been used in several FOAs to-date, including the BD2K/BISTI broad based FOAs, and the NINR Center FOA (2016) RFA-NR-16-002 (P30) and RFA-NR-16-001 (P20).

Lesson 10: Monitor compliance and uptake/use of CDEs

In addition to developing policy that encourages the use of CDEs, it can be helpful to devise measures to monitor and promote compliance and the uptake of the CDEs by the relevant communities. Some ICs withhold a portion of grant funding until investigators demonstrate that research protocols conform to designated CDEs. For example:

- The Federal Interagency Traumatic Brain Injury Research (FITBIR) Informatics System has an automated "on-boarding" process to validate that a proposed protocol conforms to designated CDEs for TBI. Only conforming protocols can be accepted.
- Within NINDS, as a term and condition of award, program staff review case report forms to verify that they make use of NINDS CDEs.
- NDAR staff monitor PubMed for published papers resulting from projects and monitor compliance with data sharing agreements.

In addition, in some databases, registries, and repositories, mechanisms such as keywords and tagging can be used to identify the use of CDEs. For example, in dbGaP, you can tag data sets that use CDEs. The rare disease registry encourages all trials using their CDEs to indicate that in the keyword field on ClinicalTrials.gov (e.g., see https://ncats.nih.gov/grdr/cdes). The caDSR repository tags those CDEs that have been identified as standards within the relevant communities. More information on these practices is available in Appendix B.

Lesson 11: Follow FAIR Data Principles in distributing CDEs

FAIR refers to "four foundational principles—Findability, Accessibility (as well as Availability and Attributability), Interoperability, and Reusability."[4]

This implies providing the CDEs in formats that are machine-readable and can be imported into data capture tools like Redcap, JSON, or XML. Avoid providing CDEs in only PDF format. After deciding upon a set of CDEs for a given initiative, the development of a data repository could be a future direction more easily enabled by the ability to import and export CDE infrastructure and collected data in machine-readable format.

Lesson 12: Harmonize Across Paradigms

There are limitations to mapping existing data sets to standard data elements which include the loss of granularity as well as many assumptions in the workflow to map the data. These difficulties are further compounded by the long-standing use of different terminologies in CDEs in varied paradigms or contexts, or when historically used labels are incompatible with current emphasis or endorsed labels. These factors can make combining and comparing data very difficult. One example is the widespread variation that has existed in collecting data historically about race, ethnicity, gender, and sexual orientation.

Additionally, traditional models of information needs, protocols, workflow, and vocabularies vary greatly between biomedical basic or clinical research, clinical care, and public health surveillance. For example, the NCI Thesaurus is used in CDISC standards, while LOINC, SNOMED CT and RxNorm are the recommended meaningful use terminologies. In addition, many EHR vendors still have and use their own proprietary data dictionaries. Issues include differences in the granularity of concepts across these terminologies, the kinds of concepts (pre- vs post-coordination), and coverage of the domain.

In such instances, it may well be that a decision must be made to go one way or the other. For example, if race, ethnicity and gender are categorized in different ways in different parts of the U. S. government, the choice as to which categorization is used to develop or select CDEs may be based on the regime in which they are most likely to be used. And, to resolve variations between CDEs for biomedical research and clinical care, a choice might be made by either having the clinical care community adopt the CDE paradigm used by some researchers or the relatively small research community adopting the standards approach used in the much larger clinical care community.

[4] Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3: 160018. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/

Appendix J: Policy Language to Encourage Use of NIH CDEs

NIH CDE Task Force | February 13, 2017

NIH Template Language

To promote the use of CDEs in NIH-funded research, NIH has added the following language to the template used for NIH Funding Opportunity Announcements (FOAs). It is included in

> *"Use of Common Data Elements in NIH-funded Research*
>
> *NIH encourages the use of common data elements (CDEs) in basic, clinical, and applied research, patient registries, and other human subject research to facilitate broader and more effective use of data and advance research across studies. CDEs are data elements that have been identified and defined for use in multiple data sets across different studies. Use of CDEs can facilitate data sharing and standardization to improve data quality and enable data integration from multiple studies and sources, including electronic health records. NIH ICs have identified CDEs for many clinical domains (e.g., neurological disease), types of studies (e.g. genome-wide association studies (GWAS)), types of outcomes (e.g., patient-reported outcomes), and patient registries (e.g., the Global Rare Diseases Patient Registry and Data Repository). NIH has established a "Common Data Element (CDE) Resource Portal" (http://cde.nih.gov/) to assist investigators in identifying NIH-supported CDEs when developing protocols, case report forms, and other instruments for data collection. The Portal provides guidance about and access to NIH-supported CDE initiatives and other tools and resources for the appropriate use of CDEs and data standards in NIH-funded research. Investigators are encouraged to consult the Portal and describe in their applications any use they will make of NIH-supported CDEs in their projects."*

IC-Specific Language

In addition to the template language, NIH ICs have developed a range of approaches for encouraging the use of CDEs by funded investigators. In some cases, a requirement or expectation to use CDEs in included in RFAs; in others, it becomes a term and condition of award. Some ICs have published notices in the NIH Guide to inform the research community of a general expectation to use CDEs in specified domains of research. Others have formulated more general policy statements to encourage the use of CDEs even beyond funded investigators. The listing below provides examples of different approaches used by NIH ICs:

Request for Applications:

1. Parkinson's Udall Centers of Excellence:

   Project and Core Directors for the Clinical Component must utilize the NINDS Common Data Elements Resource when constructing data element forms http://www.commondataelements.ninds.nih.gov. Inclusion and award of a clinical component requires compliance with collection and submission of the NINDS Core Data Elements (CDE).

2. NINDS Phase III Investigator-Initiated Multi-Site Clinical Trials:

The NINDS expects that applications submitted as an Investigator Initiated Phase III trial will use the NINDS Common Data Elements resource when constructing data collection forms. The Common Data Element website (see: http://www.commondataelements.ninds.nih.gov/) serves as a repository and dissemination tool for all NINDS CDEs for Investigators to utilize.

3. FITBIR:

   Individuals are required to comply with the instructions for the Resource Sharing Plans (Data Sharing Plan, Sharing Model Organisms, and Genome Wide Association Studies (GWAS)) as provided in the SF424 (R&R) Application Guide, with the following modifications: All applications, regardless of the amount of direct costs requested for any one year, should include a Data Sharing Plan. A data sharing plan should include use of the TBI Common Data Elements and adhere to the Federal Interagency TBI Research Informatics System (FITBIR) data policies https://fitbir.nih.gov/tbi-portal/.

4. Substance Use Disorder Data Elements

   NIDA strongly encourages investigators involved in human-subjects studies to employ a common set of tools and resources that will promote the collection of comparable data across studies and to do so by incorporating the measures from the Core and Specialty collections, which are available in the Substance Abuse and Addiction Collection of the PhenX Toolkit (www.phenxtoolkit.org). Please see NOT-DA-12-008 (http://grants.nih.gov/grants/guide/notice-files/NOT-DA-12-008.html) for further details.

5. National Institute of Nursing Research Data Elements RFA for P30 and P20 Centers:
   For the purposes of this FOA for P30 Centers, the following common data elements for symptom science must be used across center pilot studies in the self-management of symptoms in the following symptoms areas:

Topic	CDE
Pain	PROMIS-pain
Fatigue	PROMIS-fatigue
Sleep	PROMIS +additional duration question
Affective-mood	• PROMIS Positive Affect • PROMIS Depression
Affective-anxiety	PROMIS Anxiety
Affective-well being	• Psychological well-being scale • SF-36
Cognitive	PROMIS applied cognition & general concerns
Demographics	• Ethnicity • Race • Educational level • Date of birth • Gender

This list does not preclude the use of other elements that would be necessary for the outcomes of the proposed studies. In addition, all CDE data will be stored in a future NIH data repository and should be made available for sharing, if requested.

Notice of Grant Award:

- Parkinson's Disease Biomarker Program:
  A core of standardized de-identified HIPAA-compliant clinical data elements (CDEs) is available for each sample. Repository CDEs, found at http://ccr.coriell.org/Sections/Collections/NINDS/ClinicalDataForms.aspx?PgId=148&coll=ND, are broad enough for discovery research, compact enough to be streamlined, and are harmonized with the NINDS Common Data Elements: http://www.commondataelements.ninds.nih.gov/. The Repository also makes available a number of data dictionaries. Repository CDEs are developed by committee, including the NINDS, Repository staff, and experts in the field. NINDS Repository CDEs are modified as phenotypic characterizations and genetic knowledge evolve over time.....

  All samples accepted by the NINDS Repository must be collected under an IRB-approved consent that allows broad sharing of de-identified samples and clinical data for use by investigators from academia and industry for discovery research relevant to any medical disorder. ... Samples received by the NINDS Repository become the property of the NINDS, to be shared with the research community.

Guide Notices:

- Autism (NDAR)
  The widespread use of informatics resources by research communities adds significant value to research and accelerates the pace of discovery. The National Database for Autism Research (NDAR) has been established to serve the autism research community as a common platform for exchanging data, tools, and research-related information (http://ndar.nih.gov). Currently, NDAR accepts diagnostic data for a number of phenotypic assessments, genetic data, and image data in a variety of formats (for details, see http://ndar.nih.gov/ndarpublicweb/datarepository.go). Moreover, the architecture of NDAR supports confederation with other data resources (http://ndar.nih.gov/ndarpublicweb/infrastructure.go). Autism researchers who are collecting phenotypic, genetic or image data are strongly encouraged to share their data via NDAR, and to use this national resource to advance their projects. This encouragement is made to all investigators regardless of the size of the budget of their research projects, regardless of the source of support for their autism research, including investigators supported by NIH grants awarded under focused funding opportunity announcements (FOAs) or those under parent mechanism FOAs (“unsolicited” applications). Such encouragement is also extended to autism researchers supported by non-federal sources.

- NIDA
  The purpose of this Notice is to announce a major data-harmonization effort at the National Institute on Drug Abuse (NIDA), and to describe its implications for investigators in the addiction-science community. The NIDA is dedicated to advancing science by improving the yield and impact of its research portfolio. One way to accomplish this is to provide investigators with a common set of tools and resources that allow their work to span the manifold areas of the addiction sciences. Toward this end, the NIDA, in conjunction with the National Institute on

Alcohol Abuse and Alcoholism, the National Cancer Institute, the National Human Genome Research Institute, the Office of Behavioral and Social Science Research, and the broader scientific community, has identified a series of Core and Specialty measures that will promote the collection of comparable data across studies. The NIDA strongly encourages human-subject studies to incorporate the measures from the Core and Specialty collections, which are available in the Substance Abuse and Addiction Collection of the PhenX Toolkit (www.phenxtoolkit.org).

- Core: Tier 1: The measures in this collection are deemed relevant and essential to all areas of addiction science. NIDA grantees/applicants conducting human-subject studies are strongly encouraged to incorporate, at a minimum, the Core-Tier 1 measures.
- Core: Tier 2: The measures in this collection are deemed relevant to all areas of addiction science. Because Tier-2 measures are considered more burdensome and specialized than the Tier-1 measures, NIDA grantees/applicants conducting human-subject studies are strongly encouraged to incorporate them whenever possible and appropriate.
- Specialty: The measures in this collection are deemed relevant and essential within specific areas of addiction science. NIDA grantees/applicants conducting human-subject studies in the specified areas of science are strongly encouraged to incorporate the Specialty measures.

Through the use of these measures, NIDA-funded researchers will be able to share, compare, and integrate data across studies. By advocating the use of these common measures, the NIDA and its partners in science aim to further enhance knowledge about substance abuse and addiction, while advancing a culture of scientific collaboration.

Other Approaches

- NCATS (Used for establishing new patient data registries)
  Organizations who are establishing rare diseases patient registries are strongly encouraged to use the Office of Rare Disease Research/Global Rare Diseases Patient Registry Data Repository Common Data Elements (ORDR/GRDR CDEs and deposit their de-identified data into the GRDR. Although the GRDR CDEs have been developed specifically for the rare disease community, they can be used for any new patient registry. Patient Registry developers will need to consider the use of disease-specific CDEs and questions made available from NIH ICs. Registry developers who develop and utilize new CDEs and questions are encouraged to provide to NIH these CDEs and questions for consideration in the library of questions and CDEs made available by the NIH.

Appendix K: Examples of IC and non-NIH Practices for Monitoring Investigator Compliance with CDE and Data Sharing Policies

NIH CDE Task Force | February 13, 2017

FITBIR

As part of the Notice of Grant Award, grantees who are conducting clinical studies on traumatic brain injury (TBI) are expected to share their data in compliance with the Data Sharing Policy of the Federal Interagency TBI Research (FITBIR) Informatics System, and to use this national resource to advance knowledge. This is expected of all investigators regardless of the size of the budget of their research studies.

> RESTRICTION: This award is issued in the amount of $XXX,XXX total costs. Of this amount, 75% ($YYY,YYY) is restricted and may not be drawn down from the Payment Management System until the FITBIR Data Submission Request Form has been completed and uploaded. Investigators should complete the form providing assurance that all data submitted shall be in accordance with applicable polices, laws and regulations and that identities of research participants will not be disclosed. Once the requested form has been approved, the Notice of Grant Award will be revised to allow the grantee access to the remaining grant funds. See Guide Notice: http://grants.nih.gov/grants/guide/notice-files/NOT-NS-12-016.html

NINDS

Terms and Conditions of Award specify that NINDS staff must sign off on the Case Report Forms (CRFs) and use of the NINDS CDEs. The documents are reviewed to ensure they incorporate the Core CDEs where appropriate. The study staff keep the Program Director informed of any changes to their CFRs on a yearly basis and they submit a final list of the CDEs used in their dataset to NINDS. The NINDS PD and staff determine where the data should be submitted at a later date if it is already not identified in the Terms of Award.

NDAR

To monitor compliance with the requirement or expectation of data sharing, NDAR staff monitor the literature on a regular basis and create a spreadsheet to indicate when data are expected to be shared. Generally, investigators are required to deposit and share raw data every 6 months and must deposit and share derived data when a manuscript appears. NDAR staff know the NIH grant numbers associated with PIs who have data sharing expectations. When a paper citing one of those grants appears, staff compare the data in the paper with the data in NDAR. If NDAR does not have the data in the paper, staff go through a series of steps to get the data:

1. NDAR staff communicates with the PI and Data Manager.
2. NDAR staff communicates any unresponsiveness to the Program Officer who will communicate with the PI directly.
3. If no response, then the Program Officer will communicate with the institutional level signing official.
4. If still no response, the Grants Management Office will communicate with the Institutional level signing official.