CDC PLAN FOR INCREASING ACCESS TO SCIENTIFIC PUBLICATIONS AND DIGITAL SCIENTIFIC DATA GENERATED WITH CDC FUNDING

Contents

EXECUTIVE SUMMARY 4

Access to Scientific Publications 4

Access to Digital Scientific Data 5

SECTION I: BACKGROUND AND PURPOSE 8

SECTION II: ACCESS TO SCIENTIFIC PUBLICATIONS 8

A. Preamble 8

B. Scope 9

C. Requirements 9

D. Applicability 9

E. Roles and Responsibilities 10

F. Implementation 11

G. Metrics, Compliance and Evaluation 15

H. Public Consultation Experience 15

I. Interagency Coordination 16

SECTION III: ACCESS TO DIGITAL SCIENTIFIC DATA 17

A. Preamble 17

B. Planning and Existing Policy 17

C. Data Covered 18

D. Policy Development 19

E. Needs Assessment to Support Sharing of Intramural Data 21

F. Types of CDC Data to Be Shared and Mechanisms 21

G. When to Share 25

H. Existing Infrastructure Review to Determine Resource Need 26

I. Data Management 29

J. Access and Discoverability 31

K. Resources 32

L. Obligations of Extramural Researchers 34

M. Tracking Compliance 38

N. Implementation 39

SECTION IV: CONCLUSION 41

APPENDICES

EXECUTIVE SUMMARY

On February 22, 2013, the White House Office of Science and Technology Policy (OSTP) issued a memorandum titled "Increasing Access to the Results of Federally Funded Scientific Research" (OSTP Memo).[1] It directs federal agencies and offices to develop and submit plans to OSTP. The public access requirements will be applied prospectively. This plan describes the steps Centers for Disease Control and Prevention (CDC) and the Agency for Toxic Substances and Disease Registry (ATSDR) (referred to as CDC throughout this document) are taking to maximize access to intramural and extramural data without charging user fees in response to the OSTP Memo. This plan complies with OMB M-13-13, issued in May 2013.

Access to Scientific Publications

The CDC public access plan for publications outlines the requirements for the implementation of the OSTP Memo at CDC. The goal of this plan is to ensure that the public has free access to peer-reviewed publications funded by CDC. To reach this goal, authors must have a way to submit their manuscripts into a repository system that allows access to the manuscript and ensures that funded manuscripts are captured. The key points to this plan are the manuscript submission system, document repository, and compliance management.

To collect the manuscripts, CDC will use the National Institutes of Health Manuscript Submission (NIHMS) system for both intramural and extramural manuscripts. This will reduce costs as well as make submission easier for grantees that have both NIH and CDC grants.

CDC will use its CDC Stacks digital repository system for storage and access to peer-reviewed manuscripts. This system already serves as an archival repository for several agency collections. The repository meets all requirements of the OSTP Memo. The system will be upgraded to support a maximum twelve-month embargo. Peer-reviewed publications will also be dual hosted in PubMed Central to increase the distribution and archiving of CDC publications.

To ensure that all funded manuscripts are captured, both intramural and extramural compliance will be monitored. Intramural compliance will be monitored using the eClearance system (CDC system for clearing scientific information for dissemination). Extramural compliance will be monitored through Research Performance Progress Reports (RPPRs).

[1] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Access to Digital Scientific Data

The CDC public access plan for digital scientific data outlines the requirements for the implementation of the OSTP Memo and OMB M-13-13 and applies to intramural and extramural research. CDC-funded researchers will be required to make the data underlying the conclusions of peer-reviewed scientific research publications freely available in public repositories at publication in machine-readable formats. CDC will ensure data management plans include clear plans for sharing research data. CDC will also ensure new awards to researchers or institutions are not made unless the researcher has met the terms of previous awards from CDC. This includes making digital data produced in the course of previous CDC-funded research freely available, as appropriate, in compliance with the relevant data management plans for the previous awards. Program Directors will review compliance of intramural researchers through assessments of the metadata catalog and as part of the annual CIO program review.

Update the existing policy or develop new policy to ensure that data management plans are developed for all CDC research data and registered into a catalog system

The existing CDC/ATSDR Policy on Releasing and Sharing Data (CDC Data Policy) requires development of a data management plan for datasets covered by the policy. To maximize public access to digital scientific data, CDC will update its existing policy and guidelines to ensure that data management plans are developed for all such data, as required by the OSTP Memo, and establish the use of the developed generic data documentation templates (Appendices B, C, D, and E). This will help ensure metadata are captured in a database that will feed a registry to serve as a catalog of both intramural and extramural datasets. The metadata for scientific data will include, at a minimum, the common core metadata schema in use by the federal government, found at https://project-open-data.cio.gov/. Data received from public health partners are mostly surveillance data, not research data. If there are major changes in the CDC Data Policy, then CDC would discuss these changes with its partners and stakeholders.

Ensure that extramural researchers develop and provide data management plans (using CDC generic data management plan template) for data collected through federal funds to CDC and that such plans are assessed for compliance

CDC policy requires extramural research applicants to describe their “Resource Sharing Plan” (includes data) and “Translational Plan” as part of their submitted application and proposal. CDC’s data management plan template allows description of how applicants will make research data available. The data management plan will be assessed during the application and proposal review, and the quality may affect assigned scores. There will be an embargo period for researchers to publish their work before release of data in the terms of award. The solicitation/Notice of Award (NOA) will include a request for updates on the data management plan to include information regarding where data are deposited. CDC proposes to use a generic data management plan template (Appendix B) that will be electronically fillable to facilitate the capture of metadata into a catalog. The Office of Management and Budget Paperwork Reduction

Act (OMB/PRA) review and approval will be required for the implementation of the template with extramural researchers. Extramural researchers will be encouraged to use existing data standards to ensure all released data have appropriate documentation that describes the method of collection, what the data represent, and potential limitations for use.

Develop more specific guidance for determining what data to share
Data useful to scientists will be released as soon as feasible without compromising concerns about privacy, federal and state confidentiality, proprietary and national security interests, or law enforcement activities. Decisions about what data to share will take into consideration the nature of the data as it relates to issues of ownership, privacy, confidentiality, national and homeland security, dual-use research potential, trade secrets, proprietary information, and requirements of other pertinent statutes. (See sections IIIC "Data Covered" and IIIF "Types of CDC Data to Be Shared and Mechanisms").

Ensure CDC intramural data are assessed and cleared before release or sharing
Before data are released or shared, data collection, transmission, editing, processing, analysis, storage, and dissemination will be evaluated for quality and tested for completeness, validity, reliability, and ability to be reproduced. To that end, CDC will create a data review board in line with the CDC Data Policy. Decisions concerning when and how data are to be disseminated will be made on a case-by-case basis as each data product proposed for release is different. Furthermore, reviews and approval processes for releasing CDC data will vary by Centers, Institute, and Offices (CIOs) and by the type of data released.

Encourage use of public repositories and data standards
CDC will archive the research data and metadata in ways that enable preservation and access, and that also use widely available and nonproprietary formats. To achieve this, CDC will investigate available platforms for data sharing suitable for the research datasets that are produced throughout CDC and most beneficial to stakeholders. These platforms may be government or commercial, or they may use other approaches, but they will comply with OMB M-13-13 dataset requirements. All released data must have metadata required by OMB M-13-13 documentation that describes how data were collected, what the data represent, the data's completeness and accuracy, and potential limitations for use, including information to help preclude misinterpretation.

Identify and leverage technical solutions to increase access and discoverability of data
Access and discoverability enable and promote data use for a variety of users including the public, federal agencies, CDC stakeholders, and others. CDC data are shared in several ways, often in combination. Repositories can be used by CIOs that do not have or are unable to develop them. Solutions can be developed that link data registry to repositories and also have the ability to find published data in internal and external repositories. How the infrastructure solution will make data available either directly or indirectly (e.g., via a registry) in various

formats will be determined by agency-wide governance. Depending on the solution, a catalog-based registry will be developed for easier search and update of the dataset information.

SECTION I: BACKGROUND AND PURPOSE

On February 22, 2013, the White House Office of Science and Technology Policy (OSTP) issued a memorandum entitled "Increasing Access to the Results of Federally Funded Scientific Research" (OSTP Memo). The OSTP Memo directs federal agencies and offices to develop and submit plans. The public access requirements will be applied prospectively. This plan describes the steps the Centers for Disease Control and Prevention (CDC) and the Agency for Toxic Substances and Disease Registry (ATSDR) (referred to as CDC throughout this document) are taking to provide access to intramural and extramural data in response to the OSTP Memo.

CDC serves as the nation's leading agency for developing and applying disease prevention and control, environmental health, and health promotion and health education. To accomplish its mission, CDC identifies and defines preventable health problems and maintains active surveillance of diseases through epidemiologic and laboratory investigations and data collection, analysis, and distribution. In line with its role and mission, CDC analyzes and interprets data to set standards and policies that drive programs and initiatives.

In developing this plan, CDC considered input from all parts of the agency. CDC will continue to solicit feedback from its national centers and organizational units taking into consideration the understanding and agreements these units have with their stakeholders. CDC will also engage other stakeholders including libraries, publishers, federally funded researchers and universities, users of federally funded research results, and civil society groups, among others.

SECTION II: ACCESS TO SCIENTIFIC PUBLICATIONS

A. Preamble

It has long been CDC's stance that the results and accomplishments of the activities that it funds should be made available to the public. CDC believes the sharing of peer-reviewed research publications generated with CDC support will advance science and improve communication of peer-reviewed, public health-related information to scientists, public health and health care providers, policy makers, educators, and the public.

The publication public access plan is intended to: 1) create a stable archive of peer-reviewed research publications resulting from CDC-funded research to ensure the permanent preservation of these vital published research findings; 2) secure a searchable compendium of these peer-reviewed publications that CDC and its awardees can use to manage more efficiently and to better understand their research portfolios, monitor scientific productivity, and help set research priorities; and 3) make published results of CDC-funded research more readily accessible to scientists, public health and health care providers, policy makers, educators, and the public. The goal of publishing by CDC is to allow as many people as possible to read and use the science it produces. Providing free access to CDC publications increases the distribution and use of the publications. This increases the return on investment of the federal funds used to perform the research and publish the results.

B. Scope

The public access plan will be applicable to all peer-reviewed research publications funded by CDC, regardless of the funding mechanism used (e.g., grant, cooperative agreement, contract, or other funding mechanism) as well as peer-reviewed research publications authored or co-authored by CDC employees. Pursuant to the Public Access to CDC Funded Publications Policy, which was issued July 15 2013, this is applicable to all manuscripts published after this date. In its execution of the Public Access to CDC Funded Publications Policy, CDC will abide by or take into consideration law; agency mission; resource constraints; U.S. national, homeland, and economic security; and the objectives listed in the OSTP Memo.

C. Requirements

A CDC-funded author must submit an electronic version of the author's manuscript upon acceptance for publication.

1. Manuscripts resulting from extramural work must be electronically submitted to the National Institutes of Health Manuscript Submission System (NIHMS) http://www.nihms.nih.gov/.
2. Manuscripts resulting from intramural work must be electronically submitted to the NIHMS http://www.nihms.nih.gov/. The PubMed Central identification (PMCID) of the manuscript will then be entered into the CDC's eClearance system.
3. At time of submission, the submitting author must specify the date the manuscript will be publicly accessible through PubMed Central (PMC). The submitting author must also post the manuscript through PMC within 12 months of the publisher's official date of publication; however, the author is strongly encouraged to make the manuscript available earlier if possible. It is recommended that authors review the publisher's instructions to authors to determine if the publisher's required embargo time is 12 months or less. Embargo times greater than 12 months will not be allowed.

D. Applicability

The public access plan will be applicable to all peer-reviewed publications funded by CDC, regardless of the funding mechanism used (e.g., grant, cooperative agreement, contract, or other funding mechanism) as well as to all peer-reviewed publications authored or co-authored by CDC employees.

E. Roles and Responsibilities

1. Office of the Associate Director for Science (OADS)

   OADS maintains the Public Access to CDC Funded Publications Policy and IT systems required to support the policy and promote compliance. OADS manages intramural submissions to NIHMS and provides oversight for extramural submissions to NIHMS. OADS also monitors and reports compliance.
2. Centers, Institutes, and Offices (CIOs), Office of the Director Staff and Business Service Offices

   CIO Associate Directors for Science (ADS), in collaboration with OADS, are responsible for promoting compliance with the Public Access to CDC Funded Publications Policy. The Office of Science Quality (OSQ), within OADS, provides tools such as post scientific clearance compliance reports to assist ADS with ensuring manuscripts are submitted properly.
3. Procurement and Grants Office (PGO)

   PGO is responsible for including language that supports the Public Access to CDC Funded Publications Policy in funding opportunity announcements (FOAs) and requests for proposals (RFPs), and in accordance with applicable provisions found in the Department of Health and Human Services (HHS) grants regulations and policies, Office of Management and Budget circulars, and federal acquisitions regulations.
4. Supervisors

   Supervisors are responsible for ensuring compliance with the Public Access to CDC Funded Publications Policy.
5. CDC Employees

   CDC employees are responsible for uploading an electronic copy of final peer-reviewed manuscripts accepted by a journal into the NIHMS system. The PMCID of the manuscript will then be entered by the clearing author or point of contact into the eClearance system. The CDC lead author is responsible for proofing manuscripts to ensure no errors were made during the conversion in NIHMS.
6. Contractors

   In accordance with their underlying contracts, CDC contractors are responsible for submitting the peer-reviewed manuscripts into the NIHMS system and proofing manuscripts to ensure no errors were made during the conversion in NIHMS.
7. Grantees

   CDC grantees and cooperative agreement recipients are responsible for submitting the peer-reviewed manuscripts into the NIHMS system. They will then proof the manuscripts to ensure no errors were made during the conversion in NIHMS. Manuscript submissions are covered under Section 36 of OMB Circular A-110(a) and applicable HHS grants regulations.
8. Authors

If the author's meaning was altered during the proofing phase, the lead author is responsible for ensuring corrections are made to the manuscript before posting to PMC. Grammatical, formatting, readability, or copyediting changes that do not alter the meaning of the article are not required.

F. Implementation

1. Planning

The implementation of the CDC Data Plan for publications will require close coordination among CDC CIOs, external federal agencies, and private organizations.

2. Submission

Both intramural and extramural manuscripts will be electronically submitted to NIHMS.

CDC Stacks will serve as the repository for archiving manuscripts submitted through NIHMS. This will require IT infrastructure coordination with NIH as well as a funding mechanism to allow CDC to reimburse NIH for costs associated with submission of CDC-funded manuscripts.

Training on the NIHMS system will be accomplished through a combination of CDC/OADS sponsored in-person training, as well as automated online training provided by NIH.

3. Management

a. System level

i. *Infrastructure*

CDC Stacks is a full-feature repository system. The system supports full text indexing of documents in the repository, which allows for easy discovery of even minor topics. It also allows for browsing by keywords, which enables the public to narrow searches to the types of documents they are looking for.

CDC Stacks is built on an open source platform with Fedora Commons at its core. This permits a great deal of functionality and the freedom to change code without breaking a copyright with a private vendor. This also allows CDC to collaborate with public and private partners to improve the repository system and then share those improvements with the public at no cost.

CDC Stacks is indexed by all major search engines. This increases the discoverability of CDC publications and promotes use around the world.

Documents stored in CDC Stacks are stored on servers on the East Coast, West Coast, and in Atlanta, Georgia. The original scans of documents are stored for archival requirements.

CDC peer-reviewed publications will be dual-hosted in CDC Stacks and in PubMed Central. This broadens dissemination methods while inheriting the archival benefits of PubMed Central. Most documents hosted in CDC Stacks were not published in peer-reviewed publications. Adding peer-reviewed publications to CDC Stacks increases access and archival capabilities at a very minimal cost.

ii. *Hosting*
The CDC Stacks repository system is fully hosted on cloud-based servers. This reduces the cost of hosting and allows the system to scale rapidly if there is a surge in access. The reduced costs also allow for a disaster recovery site in a geographically separate zone. This helps ensure the continuity of operations for the CDC Stacks repository and acts as a backup for stored documents.

iii. *Public and Private Collaboration*
The nature of open access software encourages public and private collaboration. CDC will be working with other federal agencies that have an interest in using a similar repository system. This includes sharing code and security documentation. This will significantly reduce the effort required to bring a new repository online. Once the systems are online, code improvements can easily be shared among federal agencies.

CDC has already engaged with the Fedora Commons community. Working with the Fedora Commons project managers, CDC was able to commit CDC-generated code back to the Fedora community. These improvements are now available to anyone at no cost. CDC will continue to foster this public-private collaboration.

The Public Access to CDC Funded Publications Policy supports innovation by freely sharing all publications funded by CDC. Removing the pay wall increases the number of public and private organizations that can use CDC guidelines and best practices.

b. Item level

i. *Metadata*

The CDC Stacks repository system stores both Metadata Object Description Schema (MODS) and Dublin Core (DC) metadata for each item. The MODS schema provides flexibility, precision, and specificity when tagging documents. This enables the creation of new metadata fields tailored to specific document needs and many fields that DC does not support. CDC Stacks also supports the less detailed DC to ensure the repository is compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This ensures metadata records stored in CDC Stacks can be harvested and repurposed by other repository systems.

Metadata records for peer-reviewed publications will include attribution to authors, journals, and original publishers. The record will display the source information as well as a link to the original publisher if available.

Metadata records for peer-reviewed publications will be freely available upon first publication despite the embargo of the full text.

Both intramural and extramural researchers funded by CDC will be required to make the data underlying the conclusions presented in peer-reviewed research papers freely available in machine-readable formats at the time of initial publication unless: a) the dataset has already been made available to the public via public release or a sharing mechanism, or b) the data cannot be released due to one or more constraints listed in Section IIIC. At a minimum, the dataset release will consist of a machine-readable version of the aggregated data used in the paper's data tables. At most, the dataset release can consist of individual-level (micro) data. In cases where a minimal dataset is released with the research paper but CDC intends to release a more detailed version of the dataset after it is cleaned, documented, and vetted, the initial dataset release can be followed with a more complete data release. The second release will take place according to CDC's standard timeline for releasing research data. CDC's publications catalog will include a link to the site from which the dataset can be accessed. This provision gives CDC the flexibility to publish and share vital public health information as soon as a paper can be published; to provide the public with access to the data tables as soon as the paper is published; and to provide a more detailed, higher quality dataset within a timeframe that allows for comprehensive cleaning, documentation, and vetting of the final dataset.

In general, datasets intended for release or sharing, irrespective of publications, should be made available within 30 months of the end of data collection. This timeline applies to both intramural and extramural data.

ii. *Relationships*

The CDC Stacks repository enables the creation of relationships between documents stored in the repository and linking to external sites. The relationships, which are managed using the Resource Description Framework (RDF), allow CDC Stacks to show how documents are related. One example is the ability to link superseded CDC guidelines and recommendations to the most current version.

iii. *Licensing*

Licensing metadata will be stored for each document in the repository. This licensing metadata will be stored as part of the repository digital object and will denote what rights the public has concerning the use of the document.

c. Access and discoverability

i. Embargo

The Public Access to CDC Funded Publications Policy allows for a maximum 12-month embargo of the manuscript from publication date**.**

ii. Search

The CDC Stacks repository provides full-text indexing and keyword browsing to facilitate searches.

iii. Exposure to third-party services

The CDC Stacks repository system is compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This allows third parties to harvest the metadata records in the repository to be used in other repository systems. The ExLibris product Primo is an example of a commercial software system that is currently harvesting CDC Stacks metadata records.

CDC Stacks is designed to freely share the metadata records associated with a publication. There is no automated system for downloading publications in CDC Stacks, which limits unauthorized redistribution. Users can request copies of publications that can be freely redistributed based on publication licenses. CDC publications cohosted in PubMed Central inherit the controls developed by PMC.

iv. 508 compliance

To ensure 508 compliance, final peer-reviewed manuscripts will be converted to XML format. CDC will provide an accommodation for final publisher PDF versions that are not 508 compliant but cannot be remediated due to copyright constraints.

v. Access and discoverability

CDC Stacks is indexed by all major search engines. This increases the discoverability of CDC publications and promotes dissemination around the world. CDC Stacks ensures the public can read, download, and analyze in digital form, peer-reviewed manuscripts or published documents.

G. Metrics, Compliance and Evaluation

1. Intramural

a. Compliance

Intramural compliance will be measured using the CDC eClearance scientific clearance system. Peer-reviewed research publications published by CDC authors must first obtain clearance through the eClearance system. Once clearance is achieved, the manuscript moves into a pending publication state. After peer-review, the author will submit the manuscript to the NIHMS system then return either the NIHMSID or the PMCID to the eClearance system. This will complete all eClearance tasks. By monitoring this function of eClearance, it is apparent how many publications have been sent to the publishers. The cleared published state will tell CDC how many publications have been released. This will make monitoring the intramural compliance percentage straightforward.

b. Metrics and evaluation

The eClearance reporting system will be used to generate reports showing the current compliance level for intramural publications. These reports will be generated monthly and given to each CIO. The CIOs can also run *adhoc* reports any time.

2. Extramural

a. Compliance

Extramural compliance will be monitored through RPPRs. Grantees will be asked to report either the NIHMSID or the PMCID for any publication associated with their grant. Additional compliance systems and procedures are being developed in collaboration with NIH.

H. Public Consultation Experience

The CDC participated in the Public Access to Federally-Supported Research and Development Data and Publications meeting held by the National Academies of Science (May 14–17, 2013). Public comments supported CDC’s approach for public access.

CDC will be leveraging the HHS Citizens Petition process for continuing public engagement. This will include the process for stakeholders to petition CDC for a change in the embargo period for CDC peer-reviewed publications. The FY2014 Omnibus Appropriations Bill set the maximum embargo for publications at 12 months. Petitions for changes in the embargo period will be limited to reductions below the 12-month maximum embargo.

I. Interagency Coordination

To ensure consistency and reduce development costs, CDC will work with NIH to use the NIHMS system. This will alleviate the need for CDC to develop its own submission system. Using NIHMS also ensures that CDC grantees will use the same submission system as NIH grantees. CDC has already committed funds for making publications accessible and is working with NIH to use NIHMS.

SECTION III: ACCESS TO DIGITAL SCIENTIFIC DATA

A. Preamble

CDC believes societal interest is best served when public health data are widely available. CDC strives for open and timely release of data while maintaining standards of data quality, upholding individual privacy and confidentiality, and protecting information based on national security concerns and law enforcement investigations and activities. In addition, this plan recognizes proprietary interests, business confidential information, and intellectual property rights, and seeks to avoid negative impact on intellectual property rights, innovation, and U.S. competitiveness. In the February 2013 memo from OSTP, the administration made its position on access to scientific research data clear. To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, and economic security; and the objectives listed below; digitally formatted scientific data resulting from unclassified research supported wholly or in part by federal funding should be stored and publicly accessible to search, retrieve, and analyze.

B. Planning and Existing Policy

To develop the plan to increase access to data, CDC convened a workgroup: OADS in collaboration with Office of Public Health Scientific Services (OPHSS), and Enterprise Information Technology Portfolio Office (EITPO), with members drawn from various CIOs across CDC. The Office of General Counsel provided legal guidance.

CDC has a long history of sharing data and in 2003 established the CDC/ATSDR Policy on Releasing and Sharing Data (CDC Data Policy). Each CIO was charged with developing a system for data release and sharing in response.[2] Further information regarding the systems in place and in use by CIOs will be described in this document.

This section of the document outlines and establishes a pathway for increasing public access to digitally formatted scientific data in accordance with objectives and strategies outlined in the OSTP Memo and in line with the CDC Data Policy. This plan applies to CDC data generated or collected after completion and posting of the plan. The plan includes development of a data catalog in accordance with OMB M-13-13, which was effective as of May 2013. The plan will be carried out by and applied to all CDC CIOs that conduct or support federally funded intramural or extramural public health research. The plan does not preclude sharing other data (nonresearch) not covered by it. Such data will continue to be shared as defined by the CDC Data Policy.

[2] http://isp-v-maso-apps.cdc.gov/Policy/Doc/policy385.pdf

CDC will create guidelines and a governance structure for data matters in line with the CDC Data Policy. Also, there may be a need to have an agency-level body such as an internal steering committee on public access that will occasionally review the progress of implementation of CDC's Data Plan and recommend changes in consultation with CIOs.

C. Data Covered

For the purpose of this plan and in accordance with the OSTP Memo, "data" is defined as follows:

Digitally recorded factual material commonly accepted in the scientific community as necessary to validate research findings including datasets used to support scholarly publications, but does not include laboratory notebooks, preliminary analyses, drafts of scientific papers, plans for future research, peer-review reports, communications with colleagues, or physical objects, such as laboratory specimens.[3]

1. Public health research data are those collected or generated systematically to increase the stock of knowledge, including but not limited to epidemiology, laboratory, and environmental studies. Included are microdata and aggregated data, whether or not they lead to publication, as long as they are determined to be of use or value to the scientific community. The scope covers a broad range of data including quantitative measurements, survey and interview data, observational data, and environmental data. Increasingly, it may also include genetic data, information obtained from medical records, and potentially other data types (e.g., imaging, assays).

CDC determines whether each data collection is research or nonresearch before the start of the project. Most public health surveillance efforts are not deemed research; however, surveillance systems may be determined to be research when they involve the collection and analysis of health-related data conducted either to generate knowledge that is applicable to populations and settings other than the ones from which the data were collected or to contribute to health knowledge.[4] Similarly, a publication based on nonresearch data may be deemed a research publication if the data have been reused for a research purpose; such publications fall under the purview of the OSTP Memo.

For the purpose of this plan, the following are excluded:

1) Data shared with CDC but owned by other organizations (e.g., data provided to CDC by a managed care organization, preferred provider organizations, or technology firms for a

[3] http://clinton4.nara.gov/textonly/OMB/circulars/a110/a110.html

[4] http://aops-mas-iis.cdc.gov/Policy/Doc/policy557.pdf

specific research project). Such data may be covered by other policies or procedures that reflect pertinent laws, regulations, and agreements;

2) Data collected as part of ongoing public health monitoring (nonresearch surveillance efforts; most data provided by local health departments fall into this category), program evaluation, disease outbreak investigations, or event reporting activities;
3) Data that cannot be released due to federal and state confidentiality concerns, proprietary interests, national security interests, or law enforcement activities;
4) Data obtained under licensing or data use agreements, partner agreements, or study participant agreements that restrict the release or sharing of data; and
5) Data protected from disclosure by law.

Data excluded under this research plan, because they are considered nonresearch, will continue to be shared as defined by the CDC Data Policy.

D. Policy Development

Maximize public access without charge to digitally formatted scientific data created with federal funds while: i) protecting confidentiality and personal privacy, ii) recognizing proprietary interests, business confidential information, and intellectual property rights, and avoiding negative impact on intellectual property rights, innovation, and U.S. competitiveness, and iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden (OSTP Memo 4A)

1. <u>CDC/ATSDR 2005 Policy on Releasing and Sharing Data (Appendix A)</u>

Policy description: CDC developed the CDC/ATSDR Policy on Releasing and Sharing Data (CDC-102) in 2003 and updated it in 2005 (CDC-GA-2005-14) ("CDC Data Policy") to ensure that 1) CDC routinely provides data to its partners for appropriate public health purposes and 2) data are released or shared as soon as feasible without compromising privacy concerns, federal and state confidentiality concerns, proprietary interests, and national security interests, or law enforcement activities.

Principles: The CDC Data Policy is guided by the following principles: accountability, privacy and confidentiality, stewardship, scientific practice, efficiency, and equity. These principles are applied to the release of CDC data that are released for public use without restrictions and released to particular parties with restrictions, respectively. The CDC Data Policy also ensures that CDC is in compliance with applicable federal laws, regulations and guidelines, such as the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[5], 45 C.F.R. Parts 160 and 164, where applicable; the Freedom

[5] http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/html/PLAW-104publ191.htm

of Information Act (FOIA), 5 U.S.C. §552, as amended[6]; the Privacy Act of 1974, 5 U.S.C. §552a, as amended; OMB Circular A-130[7]; and the Department of Health and Human Services (HHS) Information Quality Guidelines.[8] There are federal laws, regulations and guidelines that apply to CDC data, which vary by data system.

Scope: The CDC Data Policy applies to data collected by CDC using federal resources, data collected for CDC by other agencies or organizations using federal resources, and data reported to CDC. The policy does not cover data shared with CDC but owned by other organizations. The CDC Data Policy provides several release and sharing options in the spirit of providing access to the extent feasible. If data cannot be shared as public use due to inability to de-identify, it may be shared under a data use agreement or at research data centers under controlled conditions (see Section IIIF, "Types of CDC Data to Be Shared and Mechanisms").

Implementation: To implement the CDC Data Policy, each CIO has set up internal guidelines including the following components: evaluating data quality, assessing the risk of disclosing private or confidential information, and developing supporting documentation (e.g., data collection methods, what the data represent, the extent of data completeness, limitations of their use, data use terms and restrictions, dataset citation guidelines, codebooks, etc.). CIOs are responsible for managing and sharing their data.

Applicability: While the CDC Data Policy covers all data collected or generated by CDC or on its behalf, this plan is limited to research data as described in the OSTP Memo. However, the guiding principles expressed in the policy appear to align with those in the OSTP Memo and CDC's plan for implementation of the OSTP Memo. CDC will maximize access without charge to applicable digitally formatted scientific data.

2. Policy revisions in response to the OSTP Memo

The CDC Data Policy recognizes proprietary interests, business confidential information, and intellectual property rights, and seeks to avoid negative impact on intellectual property rights, innovation, and U.S. competitiveness. The CDC Data Policy will undergo review to ensure it is responsive to the objectives as set forth in this OSTP Memo and OMB M-13-13. A committee has been tasked to review, identify, and implement changes to the CDC Data Policy as needed. CDC has shared and continues to share data in accordance with its policy and will release the revised policy in the timeline specified by HHS. HHS has set a due date of June 2015 for all its Operating Divisions

[6] http://www.justice.gov/oip/blog/foia-update-freedom-information-act-5-usc-sect-552-amended-public-law-no-104-231-110-stat

[7] http://www.whitehouse.gov/sites/default/files/omb/assets/omb/circulars/a130/a130trans4.pdf

[8] http://aspe.hhs.gov/infoquality/guidelines/index.shtml

(OPDIVs) to publish their policies. This plan will inform the CDC Data Policy for areas not currently covered by the CDC Data Policy or may require procedural specifications, such as establishing the use of developed generic data documentation templates that will include all the metadata required by OMB M-13-13 and ensuring registration of both intramural and extramural datasets via use of the data management plan (See Data Management section).

Because the CDC Data Policy concerns all types of data, state and local public health organizations that provide nonresearch data to CDC have a stakeholder interest in this policy. CDC has a long history of working with national, nonprofit public health partners such as the Council of State and Territorial Epidemiologists (CSTE), National Association of County and City Health Officials (NACCHO), and Association of State and Territorial Health Officials (ASTHO). These organizations and individuals will be informed of CDC's plan and revised policy through existing channels, such as ongoing monthly telephone conferences, broadcast e-mails, and institutional Web sites.

E. Needs Assessment to Support Sharing of Intramural Data

The CDC OADS and the CDC OPHSS conducted a voluntary assessment among CDC scientists and data managers in 2011. The assessment's goals were: 1) to solicit feedback from across the agency on the implementation of the CDC Data Policy; 2) to assess the landscape of data sharing practices among CDC programs; and 3) to use that information to develop tools, resources, and strategies that better assist CDC's CIOs in their data sharing efforts. Aligned with this objective was an additional aim to identify strategies for tracking the use of CDC data in order to measure CDC's larger scientific impact. Survey questions addressed knowledge and implementation of the CDC Data Policy, data sharing activities among CDC programs including; barriers to data sharing, tracking of secondary use of CDC data; and programmatic needs for tools, infrastructure, and other assistance to comply with CDC's Data Policy. Infrastructure needs were expressed in terms of additional capacity, tools, and resources to leverage data sharing processes. This plan is tailored to address the needs identified in the assessment and to fulfill these needs in line with the objectives and strategies listed in the OSTP Memo, OMB M-13-13, as well as other administration policies.

F. Types of CDC Data to Be Shared and Mechanisms

A strategy for leveraging archives and fostering public-private partnerships with scientific journals relevant to the agency's research (OSTP Memo 2A),
maximize public access without charge, to digitally formatted scientific data created with federal funds, while: iii) preserving the balance between the relative value of long-term preservation and access, and the associated cost and administrative burden (OSTP Memo 4Aiii)

Strengthen data management and release practices by creating and maintaining an enterprise data inventory and a public data listing, engaging with customers to help facilitate and prioritize data release, and clarifying roles and responsibilities for promoting efficient and effective data release practices (OMB M-13-13)
Incorporate new interoperability and openness requirements into core agency processes (OMB M-13-13)

Data to be shared will include data collected by CDC using federal resources, data collected by other agencies or organizations using CDC resources (e.g., through grants, cooperative agreements, contracts, or other funding mechanisms) and data reported to CDC that are considered research data (See Section IIIC, "Data Covered"). This includes scientific data in digital form (data to support scholarly publications) and data considered of value to the scientific community (data that may not have resulted in publication). Excluded are nonresearch data and research data that cannot be shared due to one or more of the criteria listed in Section IIIC, "Data Covered". CIOs are responsible for managing and sharing their data and as such making these determinations.

The CDC Data Policy provides that data may be made available either as:

1. Data release: Dissemination of data either for public use or through an *adhoc* request so that the data steward or designee no longer controls the data. Data that are not shared under restricted access will default to public use release.
2. Data sharing: Granting certain individuals or organizations access to data that cannot be released (e.g., that contain individually identifiable or potentially identifiable data) through the use of a special data sharing agreement that governs the use and re-release of the data and is agreed upon by CDC and the data provider if applicable. This type of data sharing is further subdivided into two categories:
    - Data sharing through a special data use agreement. These are data that cannot be released publicly, but need not always be under CDC's control. Typically this will also apply to data that if limited to such a small number of high level variables to truly de-identify, would be of little or no scientific value to any resulting analyses. In such cases data can be released through the use of a special data sharing agreement that governs the use and further release of the data. Before making the data available; however, CIOs must evaluate any requests for permission to use their confidential or private data to ensure that the data will be used for an appropriate public health purpose. To the extent possible, CDC recommends the sharing of data that cannot be released for public use with public health partners.
    - Data sharing under controlled conditions. These are data that need to remain under CDC custody at all times. These data are usually analyzed by non-CDC researchers at data enclaves located at CDC facilities. Access to data with sensitive topics, rare outcomes, and on vulnerable populations may fall in this category requiring: (a) collaborations with project investigators, (b) data use agreements with Institutional

Review Board (IRB) approval, or (c) other approval processes to ensure appropriate use of the data by those accessing the data.

CDC will disseminate data it has generated/collected, subject to limits imposed by law, resources, confidentiality, technology, and data quality. Indeed, because issues related to data release can vary from project to project, CIOs may need, and may have developed, specific data release procedures for specific projects. Hence, there is the need to develop a data management plan (see Appendix B) for each dataset.

Restricted use files may be released to secure data centers (or elsewhere) as soon as they have been prepared, reviews and approval have been obtained, and supporting documentation has been prepared and approved. Public use files will be deposited in designated repositories (determined by organization or data custodian). These public use data files will be made available as downloadable files along with accompanying documentation. Availability of such data will be indicated via the data registry/catalog.

- Data underlying publications. CDC determines whether each data collection is research or nonresearch before the start of the project. [9] Most public health surveillance efforts are not deemed research; however, surveillance systems may be determined to be research when they involve the collection and analysis of health-related data conducted either to generate knowledge applicable to populations and settings other than the ones from which the data were collected or to contribute to new knowledge about the health condition. Similarly, a publication based on nonresearch data may be deemed a research publication if the data have been re-used for a research purpose; such publications fall under the purview of the OSTP Memo.

  Data underlying research papers will be published coincident with the paper's publication unless: a) the dataset has already been made available to the public via public release or a sharing mechanism, or b) the data cannot be released due to one or more of the constraints listed in Section IIIC. At a minimum, the dataset release will consist of a machine-readable version of the aggregated data used in the paper's data tables. At most, the dataset release can consist of individual-level (micro) data. In cases where a minimal dataset is released coincident with the research paper, but CDC intends to release a more detailed dataset after it is cleaned, documented, and vetted, the initial dataset release can be followed with a more complete data release. The second release will take place according to CDC's standard timeline for release of research data. CDC's publications catalog will include a link to the site from which the dataset can be accessed. This provision allows CDC the flexibility to publish and disseminate vital public health information as soon as a paper can be published; to provide the public with access to the

[9] http://aops-mas-iis.cdc.gov/Policy/Doc/policy557.pdf

data tables as soon as the paper is published; and to provide a more detailed, higher quality dataset within a timeframe that allows for cleaning, documentation, and vetting of the final dataset.

In general, datasets intended for release or sharing, irrespective of publications, should be made available within 30 months of the end of data collection. This timeline applies to both intramural and extramural data.

1. *Activity*
   a. Develop guidance and procedures for identifying types of final research data[10] suitable for sharing. This will be done taking into consideration the nature of the data as it relates to issues of ownership, privacy, confidentiality, national security, dual-use research potential, trade secrets, proprietary information, requirements of other pertinent statutes, resource constraints, value of the data, highly influential scientific assessments, influential scientific assessments, high-profile datasets (e.g., interest from the White House, Congress, OMB, HHS, the media, industry or labor stakeholders, etc.) that are used to support standard-setting and data-use agreements.
   b. Develop guidance and procedures for when research datasets are released. Work with CIOs to delineate the considerations for when datasets will be made available pertinent to the types of CDC research datasets. These considerations may include: legal, statutory, or regulatory restrictions; size and complexity of the dataset; and potential for inadvertent release of sensitive information.

It is anticipated that CDC-wide guidance development will be initiated in FY2016 after the publishing of the revised policy set for June 2015. The development of such guidance may take 6–12 months. CDC will encourage the use of established public repositories and community-based standards.

2. *Activity*

   a. Identify, promote, and establish the use of archival methods that provide final research data and metadata in ways that: provide for long-term preservation and access to the content; use widely available and nonproprietary formats; provide

[10] Final research data are recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. This may include information in any of a variety of forms such as numerical, textual, visual (e.g., pictures, and videos), or auditory (e.g., recordings). This does not mean summary statistics or tables; rather, it means the data on which summary statistics and tables are based. For the purposes of this plan, final research data do not include laboratory notebooks, partial datasets, preliminary analyses, drafts of scientific papers, plans for future research, peer-review reports, communications with colleagues, or physical objects, such as gels or laboratory specimens.

access consistent with Section 508 of the Rehabilitation Act of 1973; and enable integration and interoperability with other federal public access archival solutions and other appropriate archives. To achieve this, CDC will investigate currently available platforms for data sharing that are potentially suitable for the variety of final research datasets that are produced throughout CDC and are most beneficial to stakeholders. These platforms may constitute federal, commercial, or other platforms or approaches. In addition, it is important to identify widely available and nonproprietary formats that will be required and other data formats that will be allowed.

G. When to Share

Maximize public access without charge, to digitally formatted scientific data created with federal funds, while: ii) recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding negative impact on intellectual property rights, innovation, and U.S. competitiveness, and iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden (OSTP Memo 4Aii and 4Aiii, OMB M-13-13)

Data will be released as soon as feasible without compromising privacy concerns, federal and state confidentiality concerns, proprietary and national security interests, or law enforcement activities. In the February 2013 memo from OSTP, the administration made its position on access to scientific research data clear. To the extent feasible and consistent with applicable law and policy; agency mission; resource constraints; U.S. national, and economic security; and the objectives listed below, digitally formatted scientific data resulting from unclassified research supported wholly or in part by federal funding should be stored and publicly accessible to search, retrieve, and analyze. For certain CDC data, i.e., those from NCHS[11], a designated statistical agency, legislation already requires the prompt dissemination of data; no changes will be made to NCHS's timelines as a result of this plan. However, for other data, CDC strives to have policies that are fair to users, regardless of organizational affiliation. In its policy revision, CDC will provide further guidance to CDC principal investigators regarding timely release of research data to the public. Data underlying research papers will be published with the paper's publication unless: a) the dataset has already been made available to the public via public release or a sharing mechanism, or b) the data cannot be released due to one or more of the constraints listed in Section IIIC. At a minimum, the dataset release will consist of a machine-readable version of the aggregated data used in the paper's data tables. At most, the dataset release can consist of individual-level (micro) data. In cases where a minimal dataset is released coincident with the research paper but CDC intends to release a more detailed version of the dataset after it is cleaned, documented, and vetted, the initial dataset release can be followed with a more

[11] http://www.cdc.gov/nchs/about/policy/data_release.htm

complete data release. The second release will take place according to CDC's standard timeline for release of research data. CDC's publications catalog will include a link to the site from which the dataset can be accessed. This provision allows CDC the flexibility to publish and share vital public health information as soon as a paper can be published; to provide the public with access to the data tables as soon as the paper is published; and to provide a more detailed, higher quality dataset within a timeframe that allows for cleaning, documentation, and vetting of the final dataset. In general, datasets, irrespective of publications, that are intended for release or sharing should be made available within 30 months following the end of data collection. This timeline applies to both intramural and extramural data. Furthermore, reviews and approval processes for releasing CDC data will vary by CIO and by the type of data released. Before data are released or shared, data will be evaluated for quality.

H. Existing Infrastructure Review to Determine Resource Need

The sharing of CDC data is diverse in nature, driven not only by the types of data and the manner in which they can be shared, but also by the fact that CIOs are responsible for managing and sharing their data. Each CIO's sharing may be influenced by different considerations (e.g., collaborator cycles, the community of practice, or the need in some cases to deposit data in external repositories). A combination of approaches may be needed to optimize archival and access. CDC will ensure that suitable repositories are identified for use by CIOs and have solutions that link the data registry to repositories (internal and external) for discoverability. External data release and sharing may occur via several systems. Currently, CDC data are located in multiple venues, such as Data.gov, GitHub, CDC WONDER, Consolidated Statistical Platform, collaborator sites (e.g., Foodborne Outbreak Online Database (a subset of NORS); PulseNet; SEDRIC; GeneBank®; SharePoint sites; and the Laboratory Response Network (LRN) Results Messenger application), WISQARS, NEISS, etc. CDC WONDER allows users to query CDC data sources, including NCHS birth and death data, but applies suppression criteria to output in order to protect confidentiality.

A database of metadata (data about the datasets) that is searchable so that researchers (internal and external) could identify points of contact, information about study design and data collected, and method for requesting access to data is desirable. EITPO, in collaboration with OADS, will determine such information technology solution.

Approach for Recommending, Implementing, and Operating the Technical Solution to Provide Public Access to CDC Research Data

The purpose of this section is to outline recommended work necessary for identifying, implementing, and operating an infrastructure solution that meets the needs of CDC and federal-wide initiatives related to open access to research data, as defined previously. The infrastructure solution is defined as all system components required to index and publish cleared research data (See Data Covered; Types of Data to Be Shared and Mechanisms sections). It should be understood that the solution is described to fulfill a goal established by the *Digital Government Strategy* from OMB M-13-13.[12] This means that how the infrastructure solution will make data available, either directly or indirectly (e.g., via a registry) in various formats to CDC, other federal entities, the public, and to scientific communities, will be determined in accordance with best practices as determined by agency-wide governance (See Policy Development section).

To identify, implement, and then sustain the operation of the infrastructure solution, the CDC team will conduct the following activities, in the seven steps captured below. These activities are aligned with the HHS Enterprise Performance Life Cycle (EPLC) framework.[13] Each of the steps is marked by one or more deliverables that are formally reviewed and accepted by a cross-cutting cadre of subject matter experts (SMEs). The oversight official creates the agency-wide governance noted above (See "Data Management" Activity 1) and ensures that decisions conform to enterprise-wide policies, regulations, and best practices.

Step 1. Develop Business Needs Statement (BNS). The BNS formally documents the high-level business need for this data sharing, collaborative environment. To fulfill the vision of the strategy to increase access to data, certain business requirements will be established such as the need to index and publish research data. With a common business (or high-level) understanding of the future state, existing or proposed projects across the federal government will be reviewed to determine if a beneficial partnership could be established (e.g., members from both or all projects that could leverage a common basis project). As an example, the National Institute of Standards and Technology (NIST) proposal titled "Data Access Reference Implementation" is an approach that should be examined and then included in the set of alternatives.

Step 2. Develop the Concept of Operations (CONOPS (draft exists)), Business Case, and Project Charter. The CONOPS will clarify, at a high level, all of the data flows, processing steps, systems, and stakeholder organizations that will form the solution. CONOPS will lead to the identification of user and data governance scenarios (See Data Management and Access and Discoverability sections) needed as part of the solution to help obtain estimates on scope and level of efforts. At this point, solution requirements with specificity at the business need level will be formed to answer the "what" of the business need.

The Business Case is the document prepared to facilitate a decision about which solution to pursue. It takes the output from step 1, after folding in the CONOPS findings, through

[12] http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government-strategy.pdf
[13] http://www.hhs.gov/ocio/eplc-lifecycle-framework.pdf

an analytic process that leads to a solution that will close the gap between the current environment and the future state or the strategic vision. The solution addresses the business need and does not identify a particular product, service, or technology.

The Project Charter will formally authorize this project and will describe the business need for the project and the solution to be delivered by the project. It provides the authority to apply up to a certain level of organizational resources to project activities.

Step 3. Document Detailed Requirements. This activity involves documenting the detailed business and technical requirements for the solutions. This will include documenting the data collection, analysis, collaboration, and reporting requirements (See Data Management and Access and Discoverability sections).

Step 4. Conduct Detailed Analysis of Alternatives. The output from step 2 is a recommended solution at the level of the business need. Increasing detail about the recommended solution is necessary to identify products, services, or technologies to implement the solution. Obtaining increasing specificity from step 3 will involve matching projects or systems within the CDC or federal government infrastructure with the category of the recommended solution. A fit-gap analysis will be included to determine the degree of match between projects or existing systems and the business need of the current project. The closest match will be compared to a *de novo* approach via a cost-benefit analysis to identify the recommended technical approach.

Step 5. Develop Solution Design. This activity involves developing a detailed solution design document, including identifying hosting options, customization, and developing system functionality. Beyond the immediate automated solution, this will also include the design of the business, data, and technical processes that will need to be in place. The design document will be the key to implementing the solution as well as communicating to various upstream and downstream business partners about what they will need to do to participate in the solution.

Step 6. Implement Solution. This activity is to implement (develop, configure, customize, or reuse) the solution for indexing and publishing research datasets available across CDC.

Step 7. Conduct User Training. This activity is to conduct user training for how to use the solution to index and publish R&D data.

While these seven steps are described sequentially, their execution may occur in parallel.

I. Data Management

A strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research (OSTP Memo Strategy 2b)
Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies (OSTP Memo 4e)
In coordination with other agencies and the private sector, support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship (OSTP Memo 4i)
Maximize public access without charge to digitally formatted scientific data created with federal funds, while: iii) preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden (OSTP Memo 4Aiii)
Collect or create information in a way that supports downstream information processing and dissemination of activities by using machine-readable and open formats, data standards, and common core extensible metadata, and by ensuring information stewardship through the use of open licenses (OMB M-13-13)

All released data must have documentation that describes the method of data collection, what the data represent, completeness and accuracy of data, and potential limitations for use, including information to preclude misinterpretation. Currently, no CIO uses a single system or centralized platform for its data. Also, there is a range of processes for clearing data for release. In some cases, de-identified public release datasets are readily available; the dataset is cleared and then open access is possible. In other cases, requests are submitted and reviewed before release, on a case-by-case basis. Some data systems require approval by all stakeholders of the dataset. In some cases, supervisory and Associate Director for Science approval via eClearance is used. Some CIOs have additional requirements, including review and approval by committees, *adhoc* review committees, or other government agencies (e.g., DHS, FBI). NCHS has policies and procedures in place for managing, storing, and sharing data.[14] With oversight from the NCHS Ethics Review Board and Confidentiality Office, each division within NCHS is responsible for managing all aspects of their data, from data collection to data dissemination. Before releasing NCHS public use files, all files are assessed for disclosure risks by the NCHS Disclosure Review Board (DRB). To streamline the public access process, CDC hopes to adopt the current model NCHS uses to manage, share, and disseminate data. With additional funding to cover operational costs, the NCHS Research Data Centers (RDC) could host restricted-use files from other CDC centers. The RDC was developed to allow researchers access to data that cannot be released as public use datasets and is available to all HHS agencies at a cost. The RDC is available to all researchers, federal or non-federal, at a cost, as long as the researchers have received approval on their research proposal.

[14] http://www.cdc.gov/nchs/about/policy/data_release.htm

In addition, CDC strives to ensure ready access to records and data that will help improve and promote the health of the American public, while complying with applicable federal laws such as the Federal Records Act.[15] How long data remains accessible needs to be determined by CIOs on a case-by-case basis based on the value of keeping the data and the adherence to objectives, responsibilities, standards, guidelines, and instructions to meet federal records management regulations, laws, and best practices for the management of electronic records (CDC Records Management Policy).

Management of data involves developing and implementing supporting processes and tools for promoting data sharing.

Activity 1. Establish a Public Access Governing Board. This will be an agency-level body, such as an internal steering committee on public access that will, from time to time, as determined necessary, review the progress of implementation of CDC's data access plan and make recommendations to the responsible official for data access and CIOs as needed.

Activity 2. Develop Data Sharing Aids. Draft templates and model language (Appendices B, C, D, and E) for a data management plan, data sharing application, data access request, and data sharing agreement have been developed and will be made electronically fillable to facilitate CIO processes and procedures.

Activity 3. Clearing Research Data for Publishing. CDC will consider the adoption and implementation of a process for clearing and then preparing research datasets for public use. Existing processes, such as those implemented by NCHS, will be examined to determine their appropriateness for data management.

Activity 4. Provide Technical Support for Preparing Datasets. Even though each CIO is responsible for preparing restricted and public use files as well as supporting documentation, common standards and tools will be developed for easier preparation of datasets for sharing. Existing references (e.g., NCHS data preparation and publishing) will be leveraged for this activity.

Activity 5. Develop Data Review and Quality Review Process. CDC will consider implementation of a process for validating datasets (e.g., content, format) for sharing. Each CIO will have an *adhoc* or standing data review board that will ensure quality, validity, and suitability of data for release.

Activity 6. Provide Training and Education for Workforce Development Related to Scientific Data Management, Analysis, Storage, Preservation, and Stewardship. All employees working on science-related activities or managing extramural research funded awards at CDC will receive training on data sharing as part of the mandatory Scientific Integrity Training Course. This will ensure CDC's workforce understands the data sharing requirements and implements them in a consistent standardized approach. Further training will be identified to institutionalize the skill and knowledge necessary for ongoing, robust use of the technical solution.

[15] http://www.law.cornell.edu/uscode/text/44/chapter-31

J. Access and Discoverability

A strategy for improving the public's ability to locate and access digital data resulting from federally funded scientific research (OSTP Memo Strategy 2b)
An approach for optimizing search, archival, and dissemination features that encourages innovation in accessibility and interoperability, while ensuring long-term stewardship of the results of federally funded research (OSTP Memo Strategy 2c)
Promote the deposit of data in publicly accessible databases, where appropriate and available (OSTP Memo 4f)
Encourage cooperation with the private sector to improve data access and compatibility, including through the formation of public-private partnerships with foundations and other research funding organizations (OSTP Memo 4g)
Develop approaches for identifying and providing appropriate attribution to scientific datasets that are made available under the plan (OSTP Memo 4h)
Provide for the assessment of long-term needs for the preservation of scientific data in fields that the agency supports and outline options for developing and sustaining repositories for scientific data in digital formats, taking into account the efforts of public and private sector entities (OSTP Memo 4j)
Build information systems to support interoperability and information accessibility (OMB M-13-13)

Access and discoverability of data facilitate and promote the use of the data for a variety of user communities including the public, federal agencies, CDC stakeholders, and others. The following activities will be taken to increase the access and discoverability of data.

Activity 1. Metadata Availability. Data released through the Public and Secure Portals will be accompanied by the necessary documentation in the form of metadata about collection procedures, completeness, and limitations. Metadata describes the content, quality, and context of the data; provides links to additional information, such as quality assurance documents and data dictionaries; and supports identification and retrieval of specific datasets.

Activity 2. Develop discovery dashboards and subscriptions-based model for easier identification and consumption of the data (online analysis, visualization, and reporting).

Activity 3. Foster scientific community involvement in ongoing collaboration in the domain.

Activity 4. Depending upon the solution, develop a catalog-based registry for easier search and update of the available dataset information. This will include the use of the developed generic data management plan template as a means of collecting metadata on both intramural and extramural research. The generic template describes the data to be produced from the research, standards for collection, preparing and sharing data, and the mechanism of sharing. Data generated from use of this template (Appendix B) will populate a CDC database of all data it generates or collects but only information on data that can be shared in one form or the other will be part of the external facing catalog.

Activity 5. Establish Partnerships. This will include: collaborating with the OMB Open Data Initiative; publishing metadata in tools leveraged through the Open Data Initiative and HHSData.gov; collaborating with other HHS OPDIVs and agencies; and synergizing efforts with other OPDIVs and agencies to increase visibility of the data. CDC is also part of the Public Health Research Data Forum.[16] This effort is exploring several solutions for discoverability and attribution for data, and CDC can leverage outcomes from this initiative. Work developed to date has included commissioned research on emerging tools for data citation. CDC will also explore mechanisms, such as NIH's proposed Data Commons, for making extramural data available.

Activity 6. Communication. Implement various communication activities in scientific meetings, conferences, and workshops where CDC promotes the Open Data Initiative by sponsoring tutorials and making presentations. Communication channels (such as the NCHS listserv, the NCHS press room, and CDC Web sites) will be leveraged for communications.

K. Resources

Identification of resources within the existing agency budget to implement the plan (OSTP Memo 2f)

CDC has already committed funds for making publications accessible and is working with NIH to use NIHMS. CDC data access planning has been driven by known or identified needs as opposed to the availability of funds. The CDC Data Plan will be implemented in this same manner using agency funds. Costs associated with the centralized infrastructure to support this public access plan will be estimated and treated as an assessment chargeable to individual CDC CIOs based on the number of publications and datasets they produce that can support public access. Direct costs to the CIOs for management and administration necessary to comply with the CDC Data Plan will be funded by their existing appropriation as well as for each respective CIO.

CDC will review the costs and benefits of maintaining data repositories and leveraging other repositories including long-term preservation to inform priority setting. It may be necessary to use a phased-in approach in this implementation, if there are financial constraints.

1. <u>Existing infrastructure</u>

CIOs will continue to use repositories that are available to them. For CIOs that do not currently use a repository, options will be presented. For example, CDC has the NCHS RDC, which supports restricted-use data. In addition, CDC WONDER houses public use datasets and is capable of hosting restricted datasets as well.

2. <u>Partnerships</u>

[16] http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030689.htm

CDC will leverage solutions identified by the Public Health Research Data Forum, an initiative that seeks to promote increased access to public health research data, of which CDC is a member. CDC will explore mechanisms, such as NIH's proposed Data Commons, for making extramural data available.

3. <u>New IT solutions</u>

Following evaluation of available systems, a determination will be made regarding the best options to increase access and discoverability of data as outlined in this plan.

L. Obligations of Extramural Researchers

A strategy for leveraging existing archives, where appropriate, and fostering public-private partnerships with scientific journals relevant to the agency's research (OSTP Memo 2a)
A plan for notifying awardees and other federally funded scientific researchers of their obligations (e.g., through guidance, conditions of awards, or regulatory changes) (OSTP Memo 2d)
An agency strategy for measuring and enforcing compliance with its plan (OSTP Memo 2e)
Identification of resources within the agency budget to implement the plan (OSTP Memo 2f);
Timeline for implementation (OSTP Memo 2g)
Ensure that extramural researchers receiving federal grants and contracts for scientific research and intramural researchers develop data-management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified (OSTP Memo 4b)
Allow the inclusion of appropriate costs for data management and access in proposals for federal funding for scientific research (OSTP Memo 4c)
Ensure appropriate evaluation of the merits of submitted data management plans (OSTP Memo 4d)
Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies (OSTP Memo 4e)
Promote the deposit of data in publicly accessible databases, where appropriate and available (OSTP Memo 4f)

Extramural research is defined by grants, cooperative agreements, or contracts that are awarded by CDC to outside institutions and other eligible applicants. For the purposes of this plan, extramural research will include both foreign and domestic awardees including institutional awards to research centers, ministries of health, state health departments, and other eligible entities that support centralized resources and facilities shared by extramural investigators conducting research. A wide range of external institutions and organizations use CDC funding to support their research projects and resource needs. CDC extramural research is unique in that awards are made to the applicant's institution and not to an individual, thus providing funding for scientists, laboratories, or other key personnel conducting extramural research under the terms and conditions of the federal award.

Extramural research contributes to increasing knowledge, improving public health interventions, accelerating the impact of CDC science, decreasing public health burdens, and improving population health domestically and globally. To further these goals, extramural research awardees are expected to report their findings and describe how they intend to release and share data generated or collected with CDC funds (see definition of data that applies) as outlined by this plan. These expectations are often set out in funding opportunity announcements (FOAs) or contract documents, and are supported by applicable grant and contract regulations, institutional policies, and applicable state, local or national laws or policies.

In addition to the data sharing requirements specified in funding mechanisms, CDC may enter into memoranda of understanding (MOU), interagency agreements (IAAs), or other agreements with entities, including foreign entities that involve data. These agreements may address unique issues or concerns with respect to data collection, sharing, and use. The CDC Data Policy enables inclusion of appropriate costs for data management in proposals for funding.

1. Legislative Authority and Regulations

Extramural research is carried out in accordance with and governed by federal laws, regulations, and policies that apply to federally funded research. Generally, these laws and policies apply to eligible foreign and domestic entities. However, extramural research conducted by foreign recipients of federal funding may raise unique issues about data collection, release, and dissemination, in light of a recipient's respective country's national or local laws and regulations. These special situations will be determined and addressed by the CIO's project officer (PO) with CDC's Procurement and Grants Office (PGO).

Generally, under HHS grants policy and the CDC Data Policy, the results and accomplishments of the activities from grants and cooperative agreements that the agency funds should be made available to the public. Awardees are expected to make the results and accomplishments of their activities available to the research community and to the public. In addition, certain HHS grants regulations provide the agency with certain rights over data first produced under a CDC award and the ability to allow others to use that data. The specific scope of CDC's rights with respect to a particular grant-supported effort should be expressed in the appropriate award documents, such as the Notice of Award (NOA). Data developed by a sub-recipient may also be subject to these requirements.

For contracts, applicable federal acquisition regulations and specified terms and conditions should be used to ensure that data collected or generated are either provided to CDC for use and dissemination or otherwise made available.

CDC strives to ensure access to records and data that will help improve and promote the health of the American public, while complying with applicable federal laws such as the Federal Records Act.[17] How long data remains accessible needs to be determined by CIOs on a case-by-case basis based on the value of keeping the data and the adherence to objectives, responsibilities, standards, guidelines, and instructions to meet federal records management regulations, laws, and best practices for the management of electronic records (CDC Records Management Policy).

[17] http://www.law.cornell.edu/uscode/text/44/chapter-31

2. Plan for Reporting

As noted above, extramural research applicants are expected under grants policy to make the results of their work available to the research community and the public at large. CDC sets out in the research FOAs or NOAs the requirements for the awardees to describe their “Resource Sharing Plan” and “Translational Plan” as part of their submitted application, which should describe how applicants will make research resources and data available for research purposes to qualified individuals within the scientific community after publication. The plans may involve data collection, analysis, and dissemination as a result of the funded research project. As noted previously, current language and guidance do not provide an agency-wide systematic approach for how a data management plan is to be provided to CDC as each CIO manages this requirement independently. Scientists seeking this data for research or related activities do so via the principal investigator (PI) of each project. There is no central repository or convenient location where extramural research data are collected and stored.

For contracts, applicable federal acquisition regulations and specified terms and conditions have not been consistently considered or used to ensure that data generated or collected are either provided to CDC for use and dissemination or otherwise made available.

For this reason the following steps will be taken to improve access to these data:

a. Revise language in research FOA template and require consideration and use of appropriate language for RFP: CDC proposes to add standard language to the research application template (SF 424) that would require applicants to include a detailed plan showing how they will collect, analyze, distribute, and make available the data collected for their proposed project. This language will also include the expectation of the awardee institution regarding compliance and the ramifications of non-compliance. This language will be added to the current language under the sections “Resource Sharing Plans” and “Translation Plans.” To standardize the development of such plans, CDC has proposed a generic data management plan template (Appendix B) that will be provided to extramural researchers in the FOA. If the applicant proposes research that would not create data files covered by this plan, the applicant will be asked to clearly state this. For contracts, appropriate terms and language will be considered and used to ensure that a data management plan is proposed with the RFP. The proposed data management plan template will also be provided.

b. Conditions of Award: CDC proposes to require that the applicant’s plan be described in clear language at the time of submission of its application, acknowledging that the plan may change during the course of the research study. The data management plan will be assessed during the proposal review and the quality may affect scores assigned to proposals. In addition, to the extent appropriate, language will be included in the applicable award documents that requests the awardee keep CDC informed of proposed changes in their resource sharing and translation plans. For contracts, language regarding a data management plan will be included in the statement of work.

c. Annual monitoring of the awardee: Funded extramural research awardee's progress with the proposed data plans will be monitored annually by the CIO's PO or appropriate POC. Monitoring will include requesting updates from the awardee, reviewing annual progress and status reports, and/or providing subject matter expertise to the awardee. The PO will report this information to the POC when requested.

d. Final report requirements: The research FOA (and other funding mechanisms), along with the NOA, state that the final report of extramural awardees include a report of the "Translation of Research Findings" and information about the publications, presentations, and media coverage resulting from the CDC-funded project and subsequent related activities. This language should be strengthened to include a provision that awardees should seek to deposit a de-identified dataset and accompanying data dictionary and other documentation relevant to use of the dataset in an established repository and inform CDC via update to their resource sharing plan (including a data management plan, which metadata is required by OMB M-13-13) within a year of fulfilling the primary purpose for which data was collected or generated. Future awards will be dependent on compliance with this requirement. Extramural researchers are responsible for ensuring the quality of data released. CDC will encourage extramural researchers via its policy and FOA to make use of existing data standards to the extent feasible in managing data and also deposit data in established repositories for archiving and preservation. Metadata of extramural research will be obtained from the completed data management plan and made available via the CDC data catalog. For contracts, awardees will be held to the requirements outlined in the deliverables.

3. Proposed Requirements

a. *Extramural Process*: The primary mechanisms to inform awardees of requirements for data sharing are the applicable funding mechanism documents (e.g., the FOA, NOA, RFP, and contracts). Specific language will be included that requires the applicant to provide a data/resource sharing plan along with a translation plan to CDC. To achieve the fullest public access to data covered by this plan, CDC will strengthen its policy and procedures to ensure data management plans will be developed and complied with by all researchers whether funded by a grant, cooperative agreement, contract, or other funding mechanism. Future awards will be dependent on compliance with this requirement. The data management plan will describe how they will provide for long-term preservation of and access to the data.

b. *CDC Data Policy*: The CDC Data Policy specifies that the costs of sharing and managing data may be included in the amount of funds requested in applications. CDC will continue to allow inclusion of costs of data management and access.

c. *Reports*: Routine reviews and reports by awardees to the CDC oversight official will help monitor the progress of data sharing. Annual reports will be prepared using the information collected through the CDC metadata catalog and using RPPRs submitted through applicable PGO award management systems.

4. Resources

Limited resources are needed to implement the monitoring and reporting related to extramural research as described above. Each CIO Extramural Research Program Office (ERPO) will designate someone to serve as the POC to collect, monitor, and submit extramural research data from each awardee. Program Directors will submit an annual report to OADS and participate in agency-wide activities related to this plan.

5. Implementation

This plan will apply to new extramurally funded eligible entities. For purposes of this plan, new extramurally funded eligible entities means any applicant responding to a published research FOA, RFA, or Program Announcement (PA) beginning with the FY2016 funding cycle. This will include entities that have been funded previously (renewals) and are applying for a continuation in response to a published FOA, RFA, or PA. This plan will not apply to applicants seeking funds under the revision mechanism (formerly a competing supplement).

6. Special Circumstances

Given the variety of state, local, and foreign governments that may be funded for extramural research activities, CDC may need to consider unique laws, regulations, and policies that apply to the sharing of data collected within the respective jurisdictions. These will need to be assessed on a case-by-case basis. Given the complex nature of data collection activities between the United States and foreign governments, CDC may also have to consider entering into MOUs, IAAs, or other agreements with these foreign entities with respect to data collections, sharing, and use.

M. Tracking Compliance

Include mechanisms to ensure that intramural and extramural researchers comply with data management plans and policies (OSTP Memo 4e)
Strengthen data management and release practices by creating and maintaining an enterprise data inventory and a public data listing, creating a process to engage with customers to help facilitate and prioritize data release, and clarifying roles and responsibilities for promoting efficient and effective data release practices (OMB M-13-13)

Data registry: CDC will develop an online registry, designed to collect metadata from data management plans (Appendix B). This online registry will include intramural and extramural research data, which will allow CDC to know what datasets exist, the repository where each dataset resides, and how the data is made accessible. Data stewards, grants management officers (GMOs), contracting officers, or POs from each CIO will monitor each research portfolio through the registry.

Each CIO, in consultation with PGO, will develop systems and procedures to ensure compliance with submission of a data management plan with each award (during review of proposals for award, at the time of issuance of a NOA, and during the submission of progress reports). Future awards will be dependent on compliance with this requirement.

A data management plan will be submitted with project proposals and be reviewed using existing project approval processes and determination processes (research and nonresearch) for intramural research. During the determination review process, the data management plan will be reviewed for completeness and compliance. Guidance will be developed for the review of plans (similar to what CDC has for research and nonresearch determinations). Data management plans will require both intramural and extramural scientists seeking funding to describe how and where they will make their data available to the public and explicitly describe how they will make the data that underlies scientific publications available for discovery, retrieval, and analysis.

Under the intramural program, existing CIO annual program review and assessments of information in the metadata catalog will be used to monitor overall compliance with data management and the CDC Data Policy developed to implement this plan, including the requirements of the OMB M-13-13 Open Data Policy. Program Directors will be responsible for carrying out a prospective review of intramural researchers' compliance with the CDC Data Policy and this data plan. .

For extramural researchers, providing data management plans is part of their award. The CIO PO in coordination with PGO will monitor each funded entity's progress and compliance. Systems and procedures will be outlined in each funding opportunity and NOA or RFA and contract to inform the awardee of the requirements for reporting data collection progress and submitting a data management plan based on the funding opportunity's specifications. Extramural compliance will be monitored through periodic and final progress reports or RPPRs and review of the metadata catalog. The PO will review and determine if the progress reports have met the stated objectives of the data sharing plans. Awardees who fail to release data in a timely fashion will be subject to procedures normally used to address lack of performance (e.g., reduction in funding, restriction of funds, or award termination). Future awards will be dependent on compliance with this requirement.

N. Implementation

Timeline for implementation (OSTP Memo 2g)

This is plan applies to data generated or collected after completion and posting of the plan. Following completion and approval of the CDC Data Plan, the final plan will be developed and implemented. To fulfill requirements of OMB M-13-13, a list of CDC major investments was submitted to OMB in spreadsheet format. A CDC online catalog is being devised to hold these and all future information about CDC data. This catalog will be dependent on use of a proposed generic data management plan template (Appendix B) (See description in sections III D, H, and J).

Ongoing activities

1. HHS revised the enterprise data inventory catalog to just the datasets associated with major investments for October 2014. This was submitted on October 17, 2014. CDC has confirmation of the completion.
2. HHS Chief Technology Officer (CTO) team has initiated a data call to validate CDC datasets (81) published in the healthdata.gov. This is currently in progress. CDC closed this by November 5, 2014.
3. HHS CTO team has already made changes to HHS Enterprise Architecture repository to serve as the department Enterprise Data Inventory catalog. Right now, it contains the datasets associated with CDC major investments, http://healthdata.gov/dataset/search?f[0]=ss_ckan_author%3ACenters%20for%20Disease%20Control%20and%20Prevention.
   This will be populated with the metadata information of the other datasets after validation.

*Future pla*ns

a. Complete business plan including evaluation of infrastructure, resource requirement, and task assignments.
b. Identify resources to begin implementation of plan FY2016. This will be at the agency level and by each CIO.
c. Develop model language for use with generic data templates by December 2015.
d. Update to CDC Data Policy and release no later than June 2015, with an effective date of October 1, 2015 (FY2016).
e. CDC-wide guidance development will be initiated in FY2016 after the publishing of the revised policy set for June 2015. The development of such guidance may take 6-12 months.
f. Any changes for extramural research will take effect FY2016 at the earliest because of fiscal award timelines. Additionally, the use of electronic generic templates that will facilitate uptake into a metadata catalog may be delayed due to OMB/PRA approval processes.
g. Implementation timeline for IT solution to increase access and discoverability will be dependent upon the solution.

h. HHS is convening a workgroup to work with OPDIVs to develop common acquisition language.
i. Program Directors will review metadata catalog to evaluate compliance annually. This compliance will be addressed in the policy revision currently underway.
j. For publications, piloting with a division at CDC will begin soon after publishing the final plan, with the hopes of moving to the rest of CDC in the subsequent months.

1. Limitations

Identification of special circumstances that prevent the agency from meeting the objectives set out in this memorandum, in whole or in part (OSTP Memo 2h)

The nature of data collection activities between governments may be complex to navigate and may need to be treated on a case-by-case basis. Further, consent forms specify the scope and inform study subjects about how their data will be used, limiting the ability to extend the use of data beyond what is specified. Informing potential subjects within future consent forms that their information might be posted to a public Web site, even in a manner that is de-identified, may reduce the willingness of subjects to participate in future studies. There may be constraints to data sharing due to statutory authorities under which the data is collected, the funding mechanisms, or sensitivities related to certain populations (e.g., Artic Investigation Program, rare birth defects, etc.).

There may be a delay in the use of a generic template for the data management plan for extramural researchers as use of such instrument will be subject to OMB/PRA; the approval processes may slow implementation. However, it is possible to still ask for a data management plan and indicate what elements it should contain pending implementation of the standardized electronic instrument that will facilitate smooth uptake of metadata into a catalog.

SECTION IV: CONCLUSION

The OSTP memorandum is very timely. Public health researchers and practitioners strive to make evidence-based decisions to help ensure that programs and policies yield maximum benefit to the public’s health. Data are critical to decision making, and technology has allowed for increasingly sophisticated ways to collect, analyze, and store information about individuals and communities. Meeting the objectives of the OSTP Memo will augment our current reinvestment in assessing and promoting data sharing practices.

Appendix A
CDC/ATSDR Policy on Releasing and Sharing Data (note: appendices to the policy are not included here, only policy content)

I. BACKGROUND

The Centers for Disease Control and Prevention (CDC)[†] and the Agency for Toxic Substances and Disease Registry (ATSDR) are the nation's principal disease prevention and health promotion agencies.[1] To fulfill their missions, these agencies must collect, manage, and interpret scientific data.

CDC believes that public health and scientific advancement are best served when data are released to, or shared with, other public health agencies, academic researchers, and appropriate private researchers in an open, timely, and appropriate way. The interests of the public—which include timely releases of data for further analysis—transcends whatever claim scientists may believe they have to ownership of data acquired or generated using federal funds. Such data are, in fact, owned by the federal government and thus belong to the citizens of the United States.

However, although CDC recognizes the value of releasing data quickly and widely, CDC also recognizes the need to maintain high standards for data quality, the need for procedures that ensure that the privacy of individuals who provide personal information is not jeopardized, and the need to protect information relevant to national security, criminal investigations, or misconduct inquiries and investigations. The goal is to have a policy on data release and sharing that balances the desire to disseminate data as broadly as possible with the need to maintain high standards and protect sensitive information.

This data release/sharing policy will also ensure that CDC is in full compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA),[2] (where applicable) the Freedom of Information Act [FOIA],[3] and the Office of Management and Budget Circular A110,[4] and the Information Quality Guidelines.

II. PURPOSE

The purpose of CDC's data release/sharing policy is to ensure that (1) CDC routinely provides data to its partners for appropriate public health purposes and (2) all data are released and/or shared as soon as feasible without compromising privacy concerns, federal and state confidentiality concerns, proprietary interests, national security interests, or law enforcement activities.

III. DATA COVERED BY THIS POLICY

This policy applies to any new data collection occurring 90 days or more following approval of this policy. Existing (previously established) data collections systems should be in compliance with this policy either within 3 years of policy approval (the cycle for surveillance and information system evaluation stipulated by the CDC Surveillance Coordination Group) or at the time of data system revisions, whichever occurs first. All data should be released as soon as feasible without compromising privacy concerns, federal and state confidentiality concerns, proprietary interests, national security interests, or law enforcement activities. Requests for data during a public health emergency will be handled on a case-by-case basis. The following data are covered by this policy:

- Data collected by CDC using federal resources.
- Data collected for CDC by other agencies or organizations (through procurement mechanisms such as grants, contracts, or cooperative agreements).
- Data reported to CDC (e.g., by a state health department).[5]

For the purpose of this policy, we use the following definitions:

CDC personnel: CDC employees, fellows, visiting scientists, and others (e.g., contractors) who are involved in designing, collecting, analyzing, reporting, or interpreting data for or on behalf of CDC.

Data: Scientific records which are as accurate and complete as possible.

Data release: Dissemination of data either for public use or through an *adhoc* request that results in the data steward no longer controlling the data.

Data sharing: Granting certain individuals or organizations access to data that contain individually identifiable information with the understanding that identifiable or potentially identifiable data cannot be re-released further unless a special data sharing agreement governs the use and re-release of the data and is agreed upon by CDC and the data providers.

For a complete list of terms used in this policy, see Appendix B.

IV. DATA NOT COVERED BY THIS POLICY

This policy does not cover data shared with CDC but owned by other organizations (e.g., data provided to CDC by a managed care organizations, preferred provider organizations, or technology firms for a specific research project). Such data may be covered by other policies or procedures that reflect pertinent laws, regulations, and agreements (such as FOIA).

V. BENEFITS OF RELEASING OR SHARING CDC DATA

- Sharing data with partners involved in collecting, analyzing, or using data will improve (1) the quality of CDC data and (2) the consistency of data across CDC.
- Sharing data will also (1) ensure that CDC scientists, contractors, awardees, and grantees are held accountable for their findings, (2) provide opportunities for study results to be validated, and (3) uncover new areas for research.[6],[7]
- Quality improves when scientists share data with partners and ask for feedback during data collection and analysis.
- Releasing or sharing data can (1) improve public health practitioners' understanding of various research methods, (2) encourage analysts from other disciplines (e.g., economists, social scientists) to examine public health questions, and (3) build trust with outside partners and the public by allowing an open critique of CDC investigations.

- U.S. states and territories have a long-standing history of voluntarily reporting individually identifiable data to CDC on incident conditions or diseases that are of public health importance.~~[8]~~ Although the electronic exchange and accumulation of data on individual cases promises public health benefits, it also creates a threat to individual privacy. The Council of State and Territorial Epidemiologists asked CDC to develop procedures that balance the need for data protection with the need to share, as broadly as possible, data collected in the interest of public health. Without such a balance, data may need to be withheld from non-CDC researchers solely to protect individual privacy.

VI. GUIDANCE FOR CIOs

In this document, CDC sets forth (1) the guiding principles to be followed when releasing/sharing data and (2) the various ways in which data can be released. Each Center/CDC organization, however, is responsible for developing specific procedures for its staff to follow. Indeed, because issues related to data release can vary from project to project, Centers/CDC organizations may need specific data release procedures for each project. For example, state and local health departments have a continuing ownership and interest in whether and how CDC re-releases data they have supplied. Custodians of such data should consult the CDC-CSTE Intergovernmental Data Release Guidelines Working Group report-+http://intranet.cdc.gov/od/ocso/ssr/drgwg.pdf which contain data release guidelines and procedures for CDC programs re-releasing state-provided data. The guidelines and procedures in the Working Group report may be useful for other data systems as well.

Guiding Principles

All CDC procedures on releasing or sharing data must be guided by the following principles.

Accountability

As a public health agency of the U.S. government, CDC is accountable to the public and to the public health community for the data it produces through research. By extension, CDC scientists are accountable for their work, and their findings are subject to independent validation. CDC scientists must conduct research with integrity; the resulting data must be of the highest possible quality; and funds must be fully accounted for.

Privacy and confidentiality

CDC recommends that, unless there is a valid public health purpose (e.g., a longitudinal study that requires record linkage), programs should not collect nor maintain identifiable data.

- Trust: Any release or sharing of public health data will acknowledge that (1) data systems are built on trust between the individuals who provide personal data and the agencies that collect those data and (2) that CDC will respect the privacy rights of individuals and others who provide personal or proprietary data. All release/sharing must be consistent with the confidentiality assurances under which the data were collected or obtained.

- Privacy Act: Identifiable data that are maintained in certain systems of records may only be released in accordance with the Privacy Act (http://www.law.cornell.edu/uscode/html/uscode05/usc_sec_05_00000552---a000-.html) which generally permits disclosing such data only with consent. However, the Privacy Act does permit data release without a subject's consent under limited conditions. One example is a release that is compatible with the purpose for which the data were collected.
- Formal confidentiality protection for research subjects: Some data collected by CDC may be given formal confidentiality protection under Sections 301(d) or 308(d) of the Public Health Service (PHS) Act. Programs that apply for such protection must make a compelling case that the information sought is so sensitive that research subjects are unlikely to provide valid data without this formal confidentiality protection.~~[‡]~~ When data have formal confidentiality protection, CDC's policy is to share those data only under conditions that are consistent with the conditions under which the data were collected. It is CDC's responsibility to ensure that inadvertent disclosure does not occur (See Appendix C).

Stewardship

CDC holds data in public trust. Good stewardship of data requires that CDC release or share data in accordance with the objectives and conditions under which the data were collected or obtained and that appropriate policies and procedures for data release be set up.[9]

Scientific practice

Before any data are released/shared, all phases of data collection, transmission, editing, processing, analysis, storage, and dissemination must be evaluated for quality.[10],[11] Preliminary data from a research project may be shared with outside partners for quality assessment but not for publication. Personnel who share data for quality assessment must follow procedures that are consistent with confidentiality agreements and other constraints.

Efficiency

Releasing data to the public and sharing data with partners is an efficient way of ensuring that data are used to their full potential, that work is not duplicated, and that funds are not spent unnecessarily.

Equity

CDC affirms the principles and practices developed to ensure impartiality and credibility of federal statistical activities.[8,[12]] CDC strives to have data release policies that are fair to all users, regardless of their organizational affiliation.

VII. HOW TO RELEASE DATA

All released data must be as complete and accurate as possible, and data must be released in accordance with the guiding principles set out in this document in one of two ways:

- Release for public use without restrictions.
- Release to particular parties with restrictions.

Restrictions can be imposed because of legal constraints or because releasing the data would risk (1) disclosing proprietary or confidential information or (2) compromising national security or law enforcement interests.

CDC recommends that data be released in the form that is closest to microdata and that still preserves confidentiality.

Release of data for public use

Data that CDC collects or holds and that can be legally released to the public should be released through a public use dataset within a year after the data are evaluated for quality and shared with any partners in data collection. Procedures for releasing public use data should be consistent with CDC's Public Health Information Network's functions and specifications.

To ensure that issues of confidentiality, proprietary use, and informed consent are addressed correctly, CIOs may choose to develop specific data release plans for each dataset. Each plan should include the following:

- A procedure to ensure that confidential information is not disclosed, for example, a list of steps to reduce this risk.~~[13],[14]~~
- A procedure to ensure that data are released in a form that does not endanger national security or compromise law enforcement activities.~~[15]~~
- A procedure to ensure that proprietary data (i.e. data owned by private organizations such as Managed Care Organizations, Preferred Provider Organizations, or technology firms) are not released inadvertently.
- Analysis plans and other documentation required by the OMB regulation on data quality.
- Instructions for non-CDC users on the appropriate use of the data.
- The date the data will be released, which should be as soon as possible after they are collected, scrutinized for errors, and validated. This release should occur no more than one year after these activities.
- The formats in which the data will be released (e.g., SAS, ASCII). For each format, give specifications (e.g., variable definitions) and information on standards for transmission.~~[16]~~

CIOs may release data without restrictions for public use through the CDC Information Center. Data may also be shared through CDC/ATSDR Scientific Data Repository and its data dissemination portal CDC WONDER (URL: http://wonder.cdc.gov/welcome.html)

- Finally, CIOs may respond to individual requests.

Data shared with restrictions

To the extent possible, CDC recommends sharing data that cannot be released for public use with public health partners. For such restricted data, special data sharing agreements must be developed. Below are two examples of how data can be shared with partners; these methods are not mutually exclusive:

- Data release under controlled conditions*:* Data that cannot be released through a public use dataset or a special-use agreement may be analyzed by appropriate non-CDC researchers at CDC-controlled data centers (e.g., the Data Center established at NCHS; see http://www.cdc.gov/rdc for a description). Alternatively, CDC may consider licensing non-CDC researchers to use certain data. Licensing would allow researchers access to identifiable data by extending legal responsibilities to those external researchers.[9] Before making the data available, however, CIOs must evaluate any requests for permission to use their confidential or private data to ensure that the data will be used for an appropriate public health purpose.

- Data release through a special-use agreement*:* Data that cannot be released publicly but that need not always be under CDC's control can be released to appropriate non-CDC researchers through a special-use agreement. Such agreements should be specific about issues related to co-authorship, reviews of findings produced through using the data, reports published about those findings, and the date the data are to be returned. All data sharing agreements should include the following:
    - Evidence that the party to whom the data are being released need the data for a legitimate public health purpose.
    - A list of restrictions on the use of the data.
    - The names of every person who will have access to the data.
    - Information on any laws pertaining to the agreement.
    - Security procedures that the non-CDC user must follow to protect the data from unauthorized use and the penalties for not following them.
    - A list of restrictions on releasing analytic results.
    - Procedures for returning the data. For an example of a set of procedures, see the CDC and ATSDR policy on data release to departing employees.~~[17]~~ , ~~[18]~~
    - Provisions that govern emergency requests for identifiable or otherwise confidential data.

An example of a special-use agreement is in the CDC/CSTE Intergovernmental Data Release Guidelines Working Group Report.[5]

VIII. IMPLEMENTATION OF CDC'S DATA-RELEASE/SHARING POLICY

Each CIO will set up procedures to ensure that CDC's policy on data release/sharing is followed. No later than 1 year after this policy is approved, CIOs should send a report on their procedures to the CDC Associate Director for Science (ADS).

One way a CIO might choose to set up procedures on data release/sharing is to authorize a data-release review board to do so. This board might report to the CIO ADS, and it might include the CIO's Information Resources Manager and stewards of relevant datasets for which the CIO is responsible. Where appropriate, subject-matter experts from the CIO should advise the board on specific data release issues.

Components of CIO procedures on data release/sharing

Each CIO must ensure that the following components are in their procedures for data release and data sharing:

An evaluation of data quality:

Evaluation of data quality must include tests for completeness, validity, reliability, and reproducibility. [11]

An evaluation of the risk of disclosing private or confidential information:

Before releasing/sharing any data, the data steward must assess the risk that personal information will be disclosed and decide whether some data need to be further de-identified.[19] For example, under the Health Insurance Portability and Accountability Act (HIPAA), 18 variables are considered identifiers, the removal of which would render the dataset de-identified. This rule, while not applicable to CDC releasing public health information, serves as a useful guide for creating de-identified data and information.[2]

Those assessing the risk that confidential information will be disclosed should recommend the statistical methods to be used for disclosure protection (e.g., suppression, random perturbations, recoding, top- or bottom-coding).[20], [21] The recommended methods should balance the risk of disclosure against the possibility that reducing the risk of disclosure will also reduce the usefulness of the data for public health practice and research.

Documentation:

All released data must have documentation that shows the conditions under which the data were collected, what the data represent, the extent of the data's completeness and accuracy, and any potential limitations on their use. Careful documentation increases the likelihood that secondary data users will interpret data correctly.

Data elements to be documented are listed in Appendix D.

CDC will develop standards for the elements needed to document data. These standards could be developed on the basis of a review of best practices for data archiving.[22],[23] Specifically, CDC standards for documentation should be compatible with those of private industry. For examples of standards, see https://hpcrd.lbl.gov/staff/olken/metrication.htm; www.fgdc.gov/standards; http://www.nbii.gov/images/uploaded/8496_1166013854464_NBII_Metadata_Standard_for_Web_Res ources_Cataloguing_Version_2.2.pdf ; www.isotc211.org; http://www.icpsr.umich.edu/DDI; or http://gcmd.gsfc.nasa.gov/Aboutus/standards.

Public release disclosure statement:

Information that will preclude misinterpretation of data should accompany all released data.

Obligations of non-CDC data users

Public use data agreements should include instructions that non-CDC data users must agree not to link data with other datasets. In addition, these agreements should include instructions to report to the CDC ADS any inadvertent discovery of the identity of any person and to make no use of that discovery.

Obligations of grantees, contractors, and partners

As of three years following approval of this policy, CDC expects researchers who are supported by CDC funding to make their data available for analysis by other public health researchers. Consequently, CDC requires that mechanisms for, and costs of, data sharing be included in contracts, cooperative agreements, and applications for grants. CDC reviewers must check whether applications for CDC funds include mechanisms for, and costs of, sharing data.

The costs of sharing or archiving data may be included in the amount of funds requested in applications for first-time or continuation funds. Applicants for CDC funds who incorporate data release into their study designs can (1) readily and economically set up procedures for protecting the identities of research subjects and (2) produce useful data with appropriate documentation. Awardees who fail to release data in a timely fashion will be subject to procedures normally used to address lack of performance (e.g., reduction in funding, restriction of funds, or grant termination).[24] Researchers who contend that the data they collect or produce are not appropriate for release must justify that contention in their applications for CDC funds.

IX. MEMORANDA OF UNDERSTANDING (MOUs) ALREADY IN PLACE

CIOs should examine the MOUs they have with other organizations or agencies to ensure that they are consistent with this data release and sharing policy and with any program-specific implementations of this policy. New MOUs should be written to ensure consistency with this policy. Any CIOs with MOUs that are inconsistent with CDC's data release policies should report that fact to the CDC ADS. Include in the report information about whatever steps have been taken to bring the MOUs into compliance with CDC's data release/sharing policy.

X. TRAINING

To ensure that this policy is followed correctly, CIOs must train their personnel in the procedures for data release/sharing. They can do so in several ways: through new Human Resources Management Office (HRMO) courses, during new employee orientation programs, at ethics certification courses, or as part of training on the CIO's local area network (LAN).

XI. CDC's COMMITMENT

CDC is committed to establishing and implementing procedures based on this policy. In addition, CDC will swiftly address any breach in the policy. Breaches consist of willful acts (e.g., deliberate disclosures that constitute scientific misconduct as defined by the Office of Research Integrity) and inadvertent disclosures (e.g., errors in judgment with no intent to do harm).

Appendix B
Data Management Plan Template for CDC Datasets (Draft)

Purpose: The purpose of this template is to assist CDC dataset custodians and extramural researchers to develop data management plans. This template is intended for use with any type of CDC and CDC-funded dataset, such as nonresearch (public health practice) data received from state health departments (such as surveillance and program data); nonresearch (public health practice) data collected by CDC (such as surveillance and emergency investigation data); and research data collected or received by CDC or CDC grantees.

Background: The Centers for Disease Control and Prevention (CDC) believes that public health and scientific advancements are best served when data are shared with other public health agencies and researchers in an open, timely and appropriate way for legitimate public health purposes. However, it is of utmost importance to insure high standards of data quality, to maintain confidentiality of individuals who provide personal information, and to protect information relevant to national security. Dataset custodians should develop data management plans that are in compliance with the CDC/ATSDR Policy on Releasing and Sharing Data (available at http://isp-v-maso-apps.cdc.gov/Policy/Doc/policy385.pdf) in addition to any policies from the relevant CIO, division, and branch. In addition, data management plans for research involving human subjects should adhere to procedures approved by relevant Institutional Review Boards, if applicable. Finally, data collected and received by CDC are federal records and are subject to federal laws and rules, as described in Appendix B of the CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data. Plans for datasets provided by states should also be consistent with the CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data (available at http://www.cste2.org/webpdfs/drgwgreport.pdf).

When to use this document: A data management plan should be developed for each dataset. This include data that will not be re-released as well as for data that will be released for unrestricted public use, under restricted controlled conditions, or through special-use data sharing agreements. Ideally, this will begin during the project planning phase and will represent a mutual understanding between CDC and the data source institution(s), if any.

Note: Elements considered essential to any plan are in black ink; additional elements that will apply to some plans are in gray ink. Red ink indicates model language (development in progress) is available for the element; this language may be used or adapted as appropriate. The elements included do not necessarily constitute an exhaustive list of all possible elements for a data management plan, so users should add elements as needed.

Data Management Plan Form (Draft)

This plan describes the anticipated use and release by CDC of the dataset named below. All CDC data management plans are required to be in compliance with the CDC/ATSDR Policy on Releasing and Sharing data, available at http://isp-v-maso-apps.cdc.gov/Policy/Doc/policy385.pdf. This plan is modifiable and does not represent a legal contract between CDC and any other entity.

Dataset Name: ____

Custodial Unit / Contact Information: ____

List the CIO/division/branch housing the dataset. List current contact person.

Study / Program Description: ____

A brief description to be included here, with reference to document or website that provides detailed information.

Memoranda of Understanding (MOU) Pertaining to Dataset: (attach)

If applicable, MOUs between CDC and other organizations with controlling interests in dataset.

Data Source(s): ____

Include all dataset provider(s), e.g. State X, Y, and Z Health Departments, or "novel data collection" by CDC / Contractor A / Research Institution B / Federal Agency C / etc.

Population Represented by Dataset: ____

Describe population represented by the data, e.g. "residents of X", "inpatients at X", "users of product X".

Type of Data: ____

Briefly describe collection type, e.g. survey, focus group, record review; whether data are at individual or aggregate level; whether data collection is one-time or ongoing.

Applicability of Public Health Service Act and Privacy Act:_________________________________

State if there is an assurance of certificate of confidentiality per Public Health Service Act 301(d) or 308(d); if Privacy Act (HIPAA) provisions apply; and type of Institutional Review Board (IRB) consent, if applicable.

Data Collection Protocol: ____

A brief description to be included here, with reference to document or website that provides detailed information.

Process for Omitting Identifying Information:
Description of what identifiers are in the database, how they will be removed, and by whom.

Data Quality Protocol (*To address issues of confidentiality protection and statistical stability)*: ____

A brief description to be included here, with reference to document or website that provides detailed information. The protocol should describe methods for these elements, which may be undertaken prior to data analysis and/or prior to re-release of the dataset:

- *data validation and error resolution*
- *removal or shielding of any proprietary information*
- *removal or shielding of sensitive information (i.e. data with dual use applicability)*
- *removal or shielding of any individually identifying information including indirect identification*

Data Retention / Disposal Plan: ____
State when and how the dataset will be archived or destroyed.

Data Analysis Plan: ____
A description of planned use of the data. Can include reference to document (e.g. Information Collection Request, Research Protocol, or other) that provides more detailed information.

Publication Plan: ____
A description of planned CDC-authored and CDC-coauthored publications, including topic, type of publication, and estimated timeline.

Allowed Uses of Stored Bio-Specimens / Need for Re-Consent: ________________________________
If there may be stored specimens, list allowed and prohibited uses (e.g. antibiotic susceptibility testing, genetic sequencing) and whether/when re-consenting of study subjects would be needed, in accordance with relevant IRB approvals and consent forms.

Dataset Release Type*:	public use	special-use data	restricted	no
Circle all that apply	dataset	sharing agreement	release	release

Dataset Release Site: ____
Planned website, research data center, or access mechanism.

Dataset Release Timeline: ____

State expected timeline from data collection to release or anticipated date of release.

Data Elements to be Released: ____

List elements to be included in public use dataset / available for data sharing / restricted release.

Dataset Release Format: ____

Specify forms of datasets, e.g. SAS, ASCII, etc.; interactive data query website; mixed mode (specify)

Dataset Release Documentation:

____

List documents provided to users, e.g. variable definitions, codebook, guidance on data use

Data Release Notification: ____

State how potential users will be informed of dataset availability.

Criteria for Data Access Eligibility:

____

For special-use and restricted releases, list criteria potential users must meet for access.

Exceptions for Emergency Needs:

____

If there are foreseeable emergencies that might require release of data under circumstances other than those described above, describe emergencies and process for emergency data release.

Date This Form Filled / Last Revised:

____

Signature of CIO/Division/Branch; Principal Investigator (extramural research); or Oversight Official:___________________________________

Signature of approving official, if required by program.

Model Element Language (in development)

Data Quality Protocol:
To address issues of confidentiality protection and statistical stability, the following procedures will be used:

- *removal of proprietary information*
- *removal of sensitive information*
- *removal of individually-identifying information (i.e. names, addresses, SSNs, medical record numbers, telephone numbers, email addresses, timing of events such as birth dates)*
- *unit limits, e.g. data will not be made public on any unit smaller than ____*
- *aggregation of data (temporally, spatially, by race, et cetera)*
- *smoothing, e.g. across geographic units*
- *suppression of data, e.g. if the total number of cases in a cell is ≤X, the cell data will be suppressed*
- *suppression of data or flagging of measures as unstable if the relative standard error of a cell is ≥30%*
- *limiting use or release to a subset of records or fields*

Appendix C
Template for CDC Public Use Dataset Release Language (Draft)

Purpose: The purpose of this template is to assist CDC dataset custodians develop information to accompany public use dataset when they are released.

Background: The Centers for Disease Control and Prevention (CDC) believes that public health and scientific advancements are best served when data are shared with other public health agencies and researchers in an open, timely and appropriate way for legitimate public health purposes. Datasets that do not contain confidential or private data can be made available for public use. Typically, public use datasets can be accessed by CDC and non-CDC researchers through web servers or other public domains. Public use dataset releases should be in compliance with the CDC/ATSDR Policy on Releasing and Sharing Data (http://isp-v-maso-apps.cdc.gov/Policy/Doc/policy385.pdf).

When to use this form: Many CDC datasets are made available to all potential users through public use datasets. When these are released, accompanying information should advise users of the location of dataset documentation, appropriate use of the data, and any restrictions on the use of the data. This information can assist users to work productively and help ensure that CDC's guiding principles of accountability, privacy and confidentiality, stewardship, scientific practice, efficiency, and equity are adhered to.

Note: In this template, elements considered essential to any public use release are in black ink; additional elements that will apply to some releases are in gray ink. Red ink indicates model language (development in progress) is available for the element; this language may be used or adapted as appropriate. Additional elements that may be needed for particular datasets can be added according to division or branch policy and program needs.

CDC Public Use Dataset Release Form (Draft)

Dataset Name: ____

CDC Custodial Unit / Contact Information: ____

List the CIO/division/branch housing the dataset. List current contact person.

Dataset Description: ____

Brief description of dataset's program purpose and contents.

Dataset Release Documentation: ____

Include documentation needed by user or reference to documentation source.

Approved Data Use and Analysis: ____

Approved uses of the data.

Restrictions on Use of Data: ____

Include all relevant restrictions imposed by CDC and program policy, law, MOUs, agreements with data sources.

Maintaining Confidentiality and Requirements if Individual Identity Discovered:

Include requirements to maintain confidentiality and notify CDC of breaches.

Requirement to Cite Data Source in Products and Publications:

If appropriate, include requirement that all oral or written presentations of results will acknowledge CDC and program providing the data, as well as source if applicable.

Citations of laws pertaining to agreement:

List any relevant state or national laws that govern the agreement and/or use of the data.

Request for notification and/or review of products and publications:

Include requests that users submit presentation slides, posters, and publications to CDC and other relevant partners for review prior to publication or presentation; and that users notify CDC and other relevant partners of acceptances and publications using the data.

Model Element Language (in development)

Approved Data Use and Analysis:
These data may be used only for the purpose of health statistical analysis and reporting.

Restrictions on Use of Data:
Any effort to determine the identity of any individual, group or organization whose information appears in the dataset is prohibited. Users may not link these data files with individually-identifiable data from other data files.

Maintaining Confidentiality and Requirements if Individual Identity Discovered:
It is of utmost importance that the identity of data subjects cannot be disclosed. All direct identifiers, as well as characteristics that might lead to identification, are omitted from the dataset. If an individual identity is discovered, make no use of the identity and immediately advise (name or position, phone number), and no one else, of this discovery.

Requirement to Cite Data Source in Products and Publications:
All written and oral presentations of results of analyses should include an acknowledgement of CDC as the source of the data.

Appendix D
Data Access Request Template for CDC Surveillance and Research Datasets (Draft)

Purpose: The purpose of this template is to assist CDC dataset custodians develop application forms for external researchers to use when requesting access to datasets, in accordance with each dataset's criteria.

Background: The Centers for Disease Control and Prevention (CDC) believes that public health and scientific advancements are best served when data are shared with other public health agencies and researchers in an open, timely and appropriate way for legitimate public health purposes. Each CDC dataset should have a data management plan that includes specification of criteria for allowing access to researchers outside the program that owns the dataset.

When to use this form: Some datasets that contain confidential or private data and are not available for public use can be accessed by CDC and non-CDC researchers through a special-use data sharing agreement, or restricted release under controlled conditions. Restricted release under controlled conditions allows access for analysis by CDC or non-CDC researchers at a CDC-controlled site (e.g. NCHS's Research Data Center) of specified data elements for a specific, approved purpose.
Before making data available through these mechanisms, CIOs must evaluate requests for data to ensure that the dataset will be used for an appropriate public health purpose and that CDC's guiding principles of accountability, privacy and confidentiality, stewardship, scientific practice, efficiency, and equity are adhered to. These principles are described in the CDC/ATSDR Policy on Releasing and Sharing Data (http://isp-v-maso-apps.cdc.gov/Policy/Doc/policy385.pdf).

Note: The precise criteria for dataset access will depend upon the characteristics of the dataset and the program; thus, dataset access request forms will need to be tailored to individual needs. However, some needs are common to all request forms. This language may be used or adapted as appropriate. Additional questions that may be needed for particular datasets can be added at each program's discretion.

Merely submitting a request does not automatically confer dataset access. Individual programs must review and approve or disapprove each request. Post approval, successful applicants will sign a data sharing agreement with CDC or CDC program that describes the roles and responsibilities of each involved agency. A special-use data sharing agreement allows release to appropriate CDC or non-CDC researchers of specified data elements for a specific, approved purpose. See Data Sharing Agreement template.

Data Access Request Form (Draft)

This form is to be completed by researchers requesting access to the CDC/ATSDR dataset named below. This is an application and not a legal contract between CDC and any other entity.

Dataset Name: ____

Applicants Who Will Have Access to
Data:____

Applicant should list all persons who will have access to data and identify the principal person responsible for the analysis and maintenance/security of the data. For each applicant, provide name, job title, research role, affiliation, address, phone, fax, and email.

Data Elements Requested: ____

Applicant should provide list of variables or tabulations, population subsets, data collection period.

Proposed Data Use and Analysis:
____

Applicant should provide a protocol that describes all planned uses of the data, including any linkage with other datasets.

Description of Expected Products and Publications from Analysis:
____

Applicant should describe all anticipated products, including presentations, publications, datasets, tools, etc.

Evidence that Data Access is Needed for Public Health Purpose:

Applicant should describe justification for access in terms of public health need.

Signature of Applicant: ____

Applicant must sign date and submit hard copy of application.

Submit this form to CDC Custodial Unit / Contact Information:
____

CDC Custodial Unit should provide submission information including contact person.

Appendix E
Data Sharing Agreement Template for CDC Surveillance and Research Datasets (Draft)

Purpose: The purpose of this template is to assist CDC dataset custodians to develop data sharing agreements.

Background: The Centers for Disease Control and Prevention (CDC) believes that public health and scientific advancements are best served when data are shared with other public health agencies and researchers in an open, timely and appropriate way and for legitimate public health purposes. A special-use data sharing agreement allows release to CDC or non-CDC researchers of specified data elements for an approved purpose, and can be used for datasets that are not available for public use.

Data sharing agreements must be in compliance with the CDC/ATSDR Policy on Releasing and Sharing Data (http://isp-v-maso-apps.cdc.gov/Policy/Doc/policy385.pdf) in addition to any applicable policies from the relevant CIO, division, and branch. A data sharing agreement will help ensure that CDC's guiding principles of accountability, privacy and confidentiality, stewardship, scientific practice, efficiency, and equity are adhered to. In addition, data use agreements for research involving human subjects should adhere to procedures approved by relevant Institutional Review Boards, if applicable. Finally, data collected and received by CDC are federal records and are subject to federal laws and rules, as described in Appendix B of the CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data. Agreements for datasets provided by states should also be consistent with the CDC-ATSDR Data Release Guidelines and Procedures for Re-release of State-Provided Data (available at http://www.cste2.org/webpdfs/drgwgreport.pdf).

When to use this form: Datasets that contain confidential or private data and are not available for public use can be accessed by CDC and non-CDC researchers through a special-use data sharing agreement. A signed data sharing agreement is a contract between CDC and the signatory data users.

These special-use agreements are to be implemented only after potential users have demonstrated a legitimate public health need and an understanding of the restrictions on the use of the data. This is accomplished through use of an application form and review process by the program. See Data Access Request Template for CDC Surveillance and Research Datasets.

Note: The precise criteria for each agreement will depend upon the characteristics of the dataset, the program, and the requirements of interested partners, so dataset agreement forms will need to be tailored to individual needs. However, some needs are common to all agreements. In this template, elements considered essential to any sharing agreement are in black ink; additional elements that will apply to some agreements are in gray ink. Red ink indicates model language (development in progress) is available for the element; this language may be used or adapted as appropriate. Additional elements that may be needed for particular datasets can be added at each program’s discretion.

Data Sharing Agreement Form for CDC Surveillance and Research Datasets (Draft)

This data sharing agreement ensures that CDC's guiding principles of accountability, privacy and confidentiality, stewardship, scientific practice, efficiency, and equity are adhered to. A signed data sharing agreement is a contract between CDC and the signatory data users.

Dataset Name: ____

CDC Custodial Unit / Contact Information:

____

List the CIO/division/branch housing the dataset. List current contact person.

Applicants Who Will Have Access to Data:____

List all persons (name, job title, research role, affiliation, email, phone) approved to have access to data and identify the principal person responsible for the analysis and maintenance/security of the data.

Project Description

Describe why this dataset is needed.

Data Elements and Format: ____

List of variables or tabulations, population subsets, data collection time period; data format.

Procedures for Provision of Data:

Describe whether data will be provided directly to the researchers, or if access will be provided to a restricted online database or at a CDC-controlled site (e.g. NCHS's Research Data Center).

Dataset Release Documentation:

____

Include documentation needed by user or reference to documentation source.

Provisional Dataset Disclaimer:

____

Include if release is not final dataset.

Period of Approval to Use Data:

____

State end date if approval is not open-ended.

Approved Data Use and Analysis:

____

Approved uses of the data, including any linkage with other datasets; may attach application, protocol.

Security Procedures to be Followed by Users:

List data security standards to be followed (e.g. encryption, passwords, locked files).

Restrictions on Use of Data:

____

Include all relevant restrictions imposed by CDC and program policy, law, MOUs, agreements with data sources.

Restrictions on Releasing Analytic Results:______________________________________

Include rules on microdata suppression, protection of identifiable information, etc.

Restrictions on Re-releasing Data:______________________________________

Restrictions against sharing data with third parties.

Procedures for Returning or Destroying Data:_________________________________

Include if applicable, e.g. if user leaves agency and/or use approval is time-limited.

Institutional Board Review:

If applicable, document reviewing IRB, approval date, approval expiration.

Maintaining Confidentiality and Requirements if Individual Identity Discovered:

Include requirements to maintain confidentiality and notify CDC of breaches.

Requirement to Report Data Accurately:

Include requirement that all reports and products will accurately reflect data.

Requirement to Cite Data Source in Products and Publications:

If appropriate, include requirement that all oral or written presentations of results will acknowledge CDC and program providing the data, as well as data source if applicable.

Requirement to Include CDC Disclaimer in Publications:

Include requirement that publications and presentations of the data include CDC disclaimer.

Requirement / Request for Copies of Draft and Final Publications:

Include requirement that publications and presentations be submitted in draft form for review and request for notification/copy of final publication.

Process for Amending Agreement:

Describe process (for both CDC and users) for amending agreement if needed.

Penalties for Violating Agreement:

State penalties to user for violating this agreement. Can include monitoring method(s).

Citations of laws pertaining to agreement:

List any relevant state or national laws that govern the agreement and/or use of the data.

Request for notification of products and publications:

Include requests that users submit slides, posters, and publications to CDC and other relevant partners for review prior to publication or presentation; and that users notify CDC and other relevant partners of acceptances and publications using the data.

Signature of Data Users: ____

All approved users must sign and date application.

Signature of CIO/Division/Branch Oversight Official:________________________________

Signature of approving official.

Model Element Language (in development)

Approved Data Use and Analysis:
I will use these data for statistical analysis and reporting as described in the attached proposal, titled ____.

Security Procedures to be Followed by Users:
I will protect the data file(s) I receive with a password and/or encryption. In addition, any temporary or permanent analysis files, such as those produced with analytic software, will be protected in the same manner(s).

I will treat all data at my worksite confidentially and keep the workstation locked when not in my use.

I will keep all hardcopies of data and analytic results locked in a secure desk or file cabinet when not in use. I will shred them when they are no longer necessary to my analysis and reports.

If I telecommute, I will follow the same procedures at my remote workstation.

I will not send or use the data with an insecure internet connection or in insecure format.

I will not produce or maintain any copies of the data.

Restrictions on Use of Data:
I will not use these data except for statistical analysis and reporting as described in the attached proposal.

Any effort to determine the identity of any individual, group or organization whose data appears in the dataset is prohibited. I will not link these data files with individually identifiable data from other data files.

Restrictions on Releasing Analytic Results:
I will not disclose or otherwise make public data on any unit smaller than _______________.

If the total number of cases in a cell is ≤X, the cell data will be suppressed in oral and written presentations.
or

If the relative standard error of a cell is ≥30%, the cell data will be suppressed in oral and written presentations.

Restrictions on Re-releasing Data:
I will not release the dataset or any part of it to any person other than those listed as collaborators in the attached proposal.

Procedures for Returning or Destroying Data:
When the proposed analyses are completed, all copies of these data will be destroyed (with confirmation in writing submitted to name or position, email address) or returned to CDC (name or position, mailing address).

Maintaining Confidentiality and Requirements if Individual Identity Discovered:
It is of utmost importance that the identity of data subjects cannot be disclosed. All direct identifiers, as well as characteristics that might lead to identification, are omitted from the dataset. If an individual identity is discovered, I will make no use of the identity and will immediately advise (name or position, phone number), and no one else, of this discovery.

Release of Coded Data
When coded data are being released, CDC will not provide the key linking codes to identifiers under any circumstances.

Requirement to Report Data Accurately:
All written and oral presentations will accurately reflect the data.

Requirement to Cite Data Source in Products and Publications:
All written and oral presentations of results of analyses will include an acknowledgement of CDC as the source of the data.

Requirement to Include CDC Disclaimer in Publications:
All written and oral presentations of results of analyses will include the following disclaimer:
"The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention."

Requirement / Request for Copies of Draft and Final Publications:

Copies of draft oral and written presentations will be submitted to the CDC program office at least 2 weeks prior to presentation or submission to a publisher so that CDC and program partners can be informed. CDC and partners may submit comments within this 2-week window. CDC reserves the right to refuse publication if …

CDC will be notified upon final publication of a product and provided with a copy and citation information.

Penalties for Violating Agreement:
I understand that if I violate this agreement, penalties may apply in accordance with CDC policies and federal law.
Compliance with this agreement will be monitored through pre-publication review of presentation products and/or verification of dataset destruction.