# Yale University Open Data Access (YODA) Project
# Public Comment Response to NIH Request for Information on Strategies for NIH Data Management, Sharing, and Citation

| | |
|---|---|
| **Submitter Name** *If submitting comments on behalf of another individual, please submit the name and function of that other individual. | The YODA Project |
| **Name of Organization** * | Yale University |
| **Type of Organization** * | University |
| **Role** * | Scientific Researcher |
| **Domain of Research Most Important to You or Your Organization** (e.g., cognitive neuroscience, infectious disease epidemiology) * | Open data and data transparency |
| **Type of Data That You Primarily Plan to Generate and Share** | |
| **Type of Data** * | Clinical |
| **Human_Non-Human** * | Human |
| **Repositories You or Your Organization Primarily Utilize** (Maximum: 250 words) | |

## Background

NIH has maintained the principle that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health (https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf). The agency has a long history and continued commitment to ensure that, to the fullest extent possible, the results of federally-funded scientific research are made available to and are useful for the general public, industry, and the scientific community (https://grants.nih.gov/policy/sharing.htm). Further, effective data sharing relies upon appropriate identification, adoption, and crediting of good data management and sharing practices, thus, NIH is adopting principles to make data "FAIR" (Findable, Accessible, Interoperable, and Reusable; http://www.nature.com/articles/sdata201618).

On February 22, 2013, the White House Office of Science and Technology Policy (OSTP) released its memorandum entitled Increasing Access to the Results of Federally Funded Scientific Research (http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf). This memorandum directs federal agencies and offices to develop plans to ensure peer-reviewed publications and digital scientific data resulting from federally-funded scientific research are accessible to the public, industry, and the scientific community to the extent feasible and consistent with applicable laws and policies. In coordination with the U.S. Department of Health and Human Services (HHS) (http://www.hhs.gov/open/public-access-guiding-principles/index.html), NIH responded to the memorandum by developing the National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research (Research https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf), released in February 2015. In order to implement the NIH Plan and move forward with ongoing commitments to the data sharing enterprise, NIH is considering priorities for data management and sharing (e.g., which data types have the greatest value for sharing, the costs and value of sharing different data types, including the long-term resource implications), and how to expand upon its 2003 Data Sharing Policy (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html).

## Yale University Open Data Access (YODA) Project
## Public Comment Response to NIH Request for Information on
## Strategies for NIH Data Management, Sharing, and Citation

Data and software citation allows important products of scientific research programs to be recognized and may enable more quantitative assessment of both effective sharing approaches and valuable data and software resources. Citation of data and software may provide additional incentives, as data and software sharing citation metrics could help to quantify these activities. Such data citation metrics would help to identify valuable data or software, to ensure that the researchers who produced them are appropriately attributed, and to facilitate broader re-use of valuable data and software by the broad research community (a list of ongoing data and software citation activities can be found in https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-015.html).

Scholarly publications typically include citations to previously published research articles where these citations provide context for the motivation of the current study and the interpretation of the results presented in the publication. Nonetheless, citations in many research articles are limited to previous publications and the concepts within them, and do not cite the specific scientific data, software tools, or workflows that underlie them. However, expectations of scholarly citation are evolving, and there is an apparent groundswell of support for data and software citation among the scientific research community.

Feedback obtained through this RFI is intended to be used to inform the development of NIH policies pertaining to the management and sharing of digital scientific data generated from NIH-supported research, including how these data and software should be cited, and other applicable NIH activities. Additionally, to support the long-term preservation of data and sustainability of repositories holding such data, NIH released the related "Request for Information (RFI): Metrics to Assess Value of Biomedical Digital Repositories"
(http://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-133.html).

### SECTION I. Data Sharing Strategy Development

NIH recognizes that many factors must be considered when determining what, when, and how data should be managed and shared. These factors include, for example, the purpose for sharing, supporting data re-use and reproducibility, maturity of the science, the infrastructure uniqueness of the data, and ethical considerations.

The NIH seeks comment on any or all of the following topics to help formulate strategic approaches to prioritizing its data management and sharing activities:

1. **The highest-priority types of data to be shared and value in sharing such data (Maximum: 250 words)**

   The Yale University Open Data Access (YODA) Project at the Yale-New Haven Center for Outcomes Research and Evaluation (CORE) fully supports and applauds the development of data sharing strategies by the NIH. To accelerate/maximize knowledge generated through NIH sponsored research, the YODA Project advocates for the sharing of de-identified individual patient-level clinical research data as one of the highest priority types of data to be shared. In addition to summary results, the availability of individual patient-level data from clinical research studies, including clinical trials and cohort studies, provides opportunities for evaluation of secondary endpoints or new research questions, validation of previously conducted effectiveness and/or safety research, and meta analyses. Repetitive data collection is reduced, minimizing study participants' time and effort as well as the cost of undertaking such research, and further maximizing the value of the NIH's research investments.

2. **The length of time these data should be made available for secondary research purposes, the appropriate means for maintaining and sustaining such data, and the long-term resource implications (Maximum: 250 words)**

The YODA Project recommends that data be shared within 12 months of study completion. Two suitable methods exist for sharing data securely. The first is through an online, publicly accessible data repository that meets accepted security criteria, such as Dryad or another form of cloud-based data storage. Alternatively, data could be shared directly to researchers on a request-by-request basis, such as through Box. In all cases, a Data Use Agreement (DUA) should be executed to ensure that external researchers employ responsible conduct with regard to the data. The DUA should detail any limitations around reuse of the data (i.e., data cannot be used for commercial or litigious purposes) and should require external researchers to commit to making no effort to re-identify patients from the data.

Ideally, data would be shared indefinitely. However, this notion is highly constrained by available resources. For sustainability, there needs to be a system in place for investigators to bear some of the cost burden, including explicit and sufficient budgets as part of the clinical research funding to cover the time, effort, and expense required for data de-identification and dissemination through data sharing initiatives, particularly for study teams with fewer discretionary resources. Funding opportunities should also be made available by government agencies and non-profit organizations to support the re-use and analysis of existing data resources.

**3. Barriers (and burdens or costs) to data stewardship and sharing, and mechanisms to overcome these barriers (Maximum: 250 words)**

There are several barriers to data stewardship and sharing, including:

- Incentives to researchers
- Preparation of metadata (including data dictionaries, syntax and/or software files)
- Version control (deciding which version/portion of the data is shared)
- Data storage (and costs – to data holders, to data accessors)
- Data security
- Data interoperability
- Curated access (time and costs)
- Informed consent (retrospective)
- Sustainability

Potential mechanisms to overcome these barriers include:

- Use of DOI for data/software to create value in citation
- Publicly report metrics on use of shared data/software

- Prepare resources on advanced preparation of metadata
- Adoption of common data models, standardized data formats, terminology
- Support the sharing of data/software at the same time as publication
- Numerous platforms are emerging that provide secure data platforms
- Planning for future expansion of availability of datasets should include consideration of proposed steps for how to include data from retrospective studies. In addition, moving forward, informed consent agreements should explicitly include data reuse as a possible future use of a research participant's data.
- Develop meaningful rewards/incentives for data sharing and penalties for not sharing
- Develop interventions to change the culture of data sharing

4. **Any other relevant issues respondents recognize as important for NIH to consider (Maximum words: 250)**

## SECTION II. Inclusion of Data and Software Citation in NIH Research Performance Progress Reports and Grant Applications

Currently, NIH grantees are required to report "other products of the research," including data, databases, and software, in section C5a of their annual RPPR submission (http://grants.nih.gov/grants/rppr/rppr_instruction_guide.pdf). However, limited guidance is available on how data, databases, and software should be reported or cited.

NIH recognizes that data and software citation indicates proof of productivity that translates to publications and patents. More thorough reporting of data and software products in the RPPR and in Competitive Grant Renewal applications may strengthen documentation of productivity and may also identify projects and investigators who most effectively share data and software.

The NIH seeks comment on any or all of the following topics:

1. **The impact of increased reporting of data and software sharing in RPPRs and competing grant applications to enrich reporting of productivity of research projects and to incentivize data sharing (Maximum words: 250)**

   We support and applaud the NIH in strengthening guidance on the citation of data, databases, and software. Increased reporting of data and software sharing in RPPRs is an effective way to incentivize data sharing, but would be strengthened if this information was also shared publicly.

2. **Important features of technical guidance for data and software citation in reports to NIH, which may include:**
   a. **Use of a Persistent Unique Identifier within the data/software citation that resolves to the data/software resource, such as a Digital Object Identifier (DOI: https://www.iso.org/obp/ui/#iso:std:iso:26324:ed-1:v1:en)(Maximum words: 250)**

We agree that the utilization of a persistent unique identifier to the data/software source would be valuable. However, version control may be more complex for living products or analytical code. Nevertheless, the ability to save and cite at key points in the project life cycle would allow for appropriate credit, accurate referencing, and reproducibility.

**b. Inclusion of a link to the data/software resource with the citation in the report (Maximum: 250 words)**

We support the inclusion of a direct link from the report citation to the data/software resource. This feature would enable expedient and accurate identification of the source data/software and support the integrity of results reported. Accordingly, there would also be a need to ensure the location of the data/software is static. It would also be valuable for the data/software to link back to the original manuscript just as the manuscript links to the data/software.

**c. Identification of the authors of the Data/Software products (Maximum: 250 words)**

We fully support the appropriate identification of data/software authors within these new data and software citations. Individuals principally responsible for the preparation and creation of data sets and software may not be the authors of the primary or secondary manuscripts; it is paramount that the appropriate credit and responsibility is attributed. However, if there is no link back to the original manuscript, manuscript authors who are not also authors of the data/software products may not be sufficiently motivated to do so.

**d. Granularity of data citations: when might citations point to an aggregation of diverse data from a single study and when might each distinct data set underlying a study be cited and reported separately (Maximum words: 250)**

We recognize the complexity of establishing uniform guidance for a multiplicity of scenarios in addressing the granularity of data citations. However, when an aggregation of diverse data from a single study results in the loss of granularity of each distinct data set, we would recommend that each underlying data set be cited and reported separately. If the aggregated data set is, in fact, simply a collection of the data sets without (significant) modification, then it would seem appropriate to utilize a singular citation of the aggregated data set.

**e. Consideration of unambiguously identifying and citing the digital repository where the data/software resource is stored and can be found and accessed (Maximum words: 250)**

We encourage avoiding any ambiguity in citing the digital repository of data/software. Direct links to the data/software are more likely to be helpful than a general link to the data repository. However, considerations should be made to ensure the data repository is still identifiable as well as sustainable.

3. **Additional routes by which NIH might strengthen and incentivize data and software sharing beyond reporting them in RPPRs and Competitive Grant Renewals applications (Maximum: 250 words)**

   The YODA Project strongly supports that the reporting of data and software sharing in RPPRs be made clearly visible in the research and healthcare policy communities through publicly accessible resources and repositories, such as through ClinicalTrials.gov or NIH RePORTER. Transparency is an important step forward in promoting the responsible and comprehensive dissemination of results of federally-funded research, and should be sufficiently publicized in order to provide a model for other organizations. Through rigorous reporting/citation policies set forth by the NIH, the availability and use of federally-funded clinical research data can be incentivized to generate new knowledge that will benefit society.

4. **Any other relevant issues respondents recognize as important for NIH to consider (Maximum: 250 words)**