Curation in the Context of the Census Curated Data Enterprise (CDE)

Draft for Comment

Sarah Nusser
Sallie Keller
Kenneth Prewitt
Steve Jost
Edward Wu
Zhengyuan Zhu
Stephanie Shipp https://orcid.org/0000-0002-2142-2136

Contact: sshipp919@gmail.com

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

January 29, 2025

Funding: This research was sponsored by the U.S. Census Bureau Agreement No. 01-21-MOU-06 and Alfred P. Sloan Foundation grant, G-2020-14002.

Citation: Nusser S, Keller S, Prewitt K, Jost S, Wu E, Zhu Z, Shipp S. (2022) Curation in the Context of the Census Curated Data Enterprise (CDE), University of Virginia, DOI: https://doi.org/10.18130/707z-gk14

.

Abstract

We consider the concept of *curation when transparency and agile reuse of data sources for the public good is paramount*, as is the case for the Curated Data Enterprise (CDE). The CDE seeks to curate and integrate existing and new data sources from the Census Bureau and external parties as part of a comprehensive information base to nimbly respond to emerging questions from policymakers and the public. The CDE's vision anchors on purpose-driven data reuse to create statistical products (e.g., data, tables, visualizations, statistical estimates, reports) in response to stakeholder objectives. This white paper outlines a perspective on curation for the Curated Data Enterprise that prioritizes transparency and agile reuse of data sources for the public good. This perspective was developed from listening sessions with stakeholders, discussions with Census Bureau staff, and information from scholarly literature.

DRAFT for Comment

January 29, 2025

Curation in the Context of the Census Curated Data Enterprise (CDE)
Sarah Nusser, Sallie Keller, Kenneth Prewitt, Steve Jost, Chris Barrett, Joe Salvo, Stephanie Shipp, Matthew Snipp, John Thompson, Edward Wu, Zhengyuan Zhu

A New Vision for Curation

In this white paper, we consider the concept of *curation when transparency and agile reuse of data sources for the public good is paramount*, as is the case for the Curated Data Enterprise (CDE). The CDE seeks to curate and integrate existing and new data sources from the Census Bureau and external parties as part of a comprehensive information base to nimbly respond to emerging questions from policymakers and the public. The CDE's vision anchors on purpose-driven data reuse to create statistical products (e.g., data, tables, visualizations, statistical estimates, reports) in response to stakeholder objectives (Keller et al. 2022).

This white paper outlines a perspective on curation for the Curated Data Enterprise that prioritizes transparency and agile reuse of data sources for the public good. This perspective was developed from listening sessions with stakeholders, discussions with Census Bureau staff, and information from scholarly literature.

The National Academies of Sciences, Engineering, and Medicine (2021) highlights the importance of *transparency* for official statistics for endeavors such as the CDE. On page 2 of their report, they note that "the goal of transparency is to enable consumers of federal statistics to understand and evaluate how estimates are generated accurately." Understanding and evaluating statistical products requires documentation on how data were acquired, evaluated, edited, processed, and analyzed to generate the product. These types of documentation are also necessary to facilitate the reuse of statistical products by others (Faniel et al. 2019, Nusser 2023), such as statistical agencies and their external users.

This form of transparency provides a new lens for *data curation*. Ideally, digital curation involves maintaining, preserving, and adding value to digital data throughout its lifecycle (Poole 2016). In practice, data curation has traditionally focused on employing metadata to describe formats and contents of data files and identifying the appropriate methods and specifications for making data accessible to others. When transparency and flexible data reuse are key elements of an information system such as the CDE, curation activities must be expanded beyond traditional metadata describing the statistical product to include documentation on the purpose for creating the statistical product, the methods used to create the product, and contextual information that supports use and reuse for a different purpose.

The CDE framework (Figure 1) incorporates this expanded concept of curation by explicitly calling for *purposes and uses* associated with assets to be curated. This approach enables data and statistical models to be responsibly and effectively shared with users, maximizes the potential that products are appropriately used and reused to produce new statistical information

for the public good, and reduces burden by reusing existing data and statistical products and avoiding new data collection. Objects beyond traditional metadata describing the data include, but are not limited to:

- Using the data source to produce new statistical outputs (e.g., a new data source from integrating several sources; statistical summaries such as tables and reports);
- How the data source came into being (its provenance);
- Ethical and equity considerations in using the data source and producing the statistical outputs;
- Rationales for the decisions made in generating, evaluating, processing and using the data source to produce statistical outputs; and
- Software code and other resources (e.g., paradata, instruments) that describe how data were generated, evaluated, processed, and analyzed to produce statistical outputs.

Figure 1. The Curated Data Enterprise Framework.

Curation occurs at each step of the CDE framework. The framework starts with the purpose and use framed as a research question, or problem identification, and continues through the following steps as outlined on the inner loop: subject matter input (literature reviews and discussions with experts); data discovery (inventory, screening, and acquisition); data ingestion and governance; data wrangling (data profiling, data preparation and linkage, and data exploration); fitness-for-purpose assessment; and statistics development. The outer loop includes communication, stakeholder engagement, data curation, equity and ethics reviews, privacy and confidentiality assessment, and dissemination of results. The process is iterative and not necessarily sequential.

Many potential uses of CDE resources will involve *integrating and comparing multiple data sources* to produce a new statistical output. It is illustrative to consider the concept of art curation in understanding how the curation of related data sources can be viewed. Art curators engage in two major tasks as part of their curation process. The first is to select works for quality and significance, the key to both ensuring that the exhibit honors verified works of art and that fakes

are exposed. The second task is to juxtapose two or more artistic expressions and blend them to inspire new insights by the viewer. In the context of curating data, it is essential to verify the source and evaluate its salience and quality for the purpose being addressed, eschewing sources that are of poor quality or not fit for the purpose, or applying improvements to the data source so that it can be used for the purpose. In addition, the need to triangulate (juxtapose) and integrate sources of information is required to produce novel insights to serve the purpose of the statistical analysis. Poole (2016) and others echo this view by noting the importance of adding value in digital curation.

Defining CDE Curation

To formalize these ideas, we propose a curation definition that serves the CDE's vision and framework. This definition has evolved from numerous discussions with stakeholders via listening sessions, Census Bureau discussions, and scholarly literature.

Curation involves documenting, for each CDE product, the *inputs* from which the product is derived, the *manipulations* used to transform from inputs to product, and the *CDE statistical product* itself. Each CDE product is *defined by purpose and use*.

The *CDE statistical product* may be a data source to be created, a new data source to be ingested, or a set of statistical outputs (e.g., new data source, tables/visuals, reports) that provide a response to a policy or research question.

The *purpose* describes why a statistical product(s) is being created and provides insight into perspectives used to select information sources and methodologies for creating the product.

Inputs include the *data and information* used to create the product, the *evaluations* (e.g., for quality and fitness for purpose) of the product, and *decisions* used to create the product (e.g., verifying the credibility and potential replicability of data sources; selecting and understanding methods applied in creating the product). Inputs should be maintained as versioned files and documents and may be associated with software code and outputs used to support the choice of inputs.

Documentation for *manipulations* applied creating the CDE product should be detailed. In particular, all manipulations and evaluations must be defined by versioned software (with associated documentation justifying each manipulation), so that the manipulations can be both understood and repeated, e.g., to produce updated outputs if input data changes.

Curation is viewed as a *constantly updated process* (see Dempsey et al. 2022). That is, documentation for transparency continues to be developed and amended throughout the creation of the product. This includes updating documentation as new or updated versions of the inputs, software code, or product are produced or the inputs/processes are adapted to apply to new purposes. When properly designed, a significant benefit of this approach is it is straightforward to generate snapshots of the current version of the product and associated digital materials for sharing with others, and it provides the basis for ongoing documentation of manipulations and statistical outputs as they are repurposed.

Given the importance of transparency and data reusability to the CDE, objects (such as data, metadata, code, and documents) stored in the CDE should be constructed to adhere to principles that promote open science to the extent feasible. For example, the FAIR principles (go-fair.org) call for data and associated objects to be *findable, accessible, interoperable, and reusable* by others. Nusser (2023) notes that the FAIR principles have been widely accepted since their introduction in 2016, including by the National Institutes of Health and the European Research Council. The principles prioritize machine actionability and reusability of shared objects to support transparency and clarity needed for proper understanding and reuse of data and other digital objects. The FAIR principles do not imply unrestricted access to all parts of the CDE. Rather it means that digital objects describing the existence of the product and processes used to generate the data or statistical product can be found and accessed along with descriptions of whether or how the data may be accessed.

While there is increasing awareness of the need for deeper information about data or statistical products and the processes used to generate them, *assessments of data quality and fitness for purpose* have received limited attention in broader scientific transparency discussions (Nusser 2023). However, these assessments are critical to verifying the data, understanding their quality, developing plans to address flaws, and considering ethics and equity in the new data or product context. Correspondingly, in the context of the CDE, evaluations of the quality and fitness for purpose of the data and other resources are an essential element of the rigorous and purpose-driven approach that undergirds the CDE vision. Evaluation methods and output should be curated as part of the CDE framework.

Census Bureau draft definition for data curation

The Census Bureau is embracing the need for end-to-end curation of data products, moving beyond their traditional data curation efforts. This will become central to their modernization and enterprise initiatives. In Fall 2022, a Census Bureau working group crafted a draft definition for data curation:

> Data curation encompasses efforts to support preserving and adding value to data, including:
> 1) organizing data to facilitate discovery and provide access;
> 2) documenting data to enable the reuse of the data in scientific and programmatic research; and
> 3) enhancing the value of the data ecosystem a) through linkages between datasets and b) by mapping the network of interconnections between datasets, research outputs, researchers, and institutions

This definition provides a useful perspective on the types of goals for establishing a curation process, as well as some of the activities that would need to be established in curating a data asset or a collection of data sources and tracking downstream uses of data and the statistical outputs produced. Note that the four components of FAIR data (findable, accessible, interoperable, reusable) are incorporated in this description.

Listening Session Input for Implementing CDE Curation

What would establish an expanded data curation approach for the CDE? Considering this question, it is useful to examine both the data/statistical summary production perspective and context for users to effectively take advantage of available statistical products. Producing statistical products involves generating shared data sources and statistical outputs. It also relies on tasks such as acquiring and integrating data, evaluating and amending the quality of the data, processing and analyzing the data, and sharing data with documentation and statistical products. Users obtain the statistical products and depend heavily on the quality of data and documentation shared by those who generate the products. Users engage in tasks to find and access potential data sources, evaluate and select candidate data sources, understand *the data and appropriate uses, process and analyze the data, and share statistical outputs and* possibly a new data source based on their work (Nusser 2023, Mikytuck et al. 2022).

We conducted listening sessions with experts and stakeholders regarding various aspects involved in planning and implementing the CDE. Based on these listening sessions, discussions with Census Bureau staff, and a literature review, the following material summarizes 14 topics for further exploration in developing a CDE curation approach that prioritizes transparency and agile reuse.

Producing Statistical Outputs

Considerable effort is required to prepare well-curated data for sharing using traditional approaches. It is possible that the expanded approach to CDE curation will increase that effort, at least initially. How can an implementation approach be designed to promote quality curation and reduce workload in the process? Our discussions highlighted both strategies and challenges for addressing this question.

Several ideas from our discussions and listening session on curation centered on how a CDE curation approach could promote quality and reduce burden. They include:

1. Start by developing a plan for capturing decisions made, methods used, literature reviewed, and outputs such as documents, code, and metadata, that occurs at each step in the data production process.
2. Automate data generation and document capture processes as much as possible, using tools that enable sharing and versioning during production.
3. Use or develop standards that support harmonizing data elements and formats across data sources.
4. Explore the idea of "light" curation to create flexibility in finding a balance between the quality of the curated product and the work required to create it.

To capitalize on these recommendations, developing high-level models for curation that identify functionalities and workflows for consideration in the planning phase would be useful. A model that identifies objects to be addressed in planning will ideally yield the kind of information needed to develop automated workflows (National Academies of Sciences, Engineering, and Medicine 2022) and approaches to versioned capture of methodologies and standards to support the producer team and to be shared with released statistical outputs (Dempsey et al. 2022). A complete planning model could illuminate areas of the model to pilot test for further refining the approach. In addition, alternative methodologies could be explored for balancing effort and

documentation quality, such as using machine learning approaches to generate metadata automatically.

Our discussions also offered several challenges that will need to be explored in developing curation strategies, including understanding approaches to:

5. Implement FAIR principles to the degree possible.
6. Characterize data quality to evaluate a statistical product or develop documentation on appropriate uses and potential flaws.
7. Define the role of data providers (e.g., human subjects) with respect to future uses or as a source for understanding data content and quality.
8. Partner with organizations to obtain external sources, such as negotiating terms of acquisition, understanding and defining permissions for downstream use, and crafting a curation process/partnership for ingesting the data source.
9. Track provenance for multiple or dynamic data sources.

Any one of these topics could be the basis for a research initiative. For example, it would be useful to explore how FAIR principles are implemented by others, what works and what is more problematic, and how this applies in an official statistics context. Again, this kind of exploration could highlight areas that are more readily adopted to begin the process of supporting FAIR digital objects in the CDE.

User Perspectives

Whether internal or external to the agency producing statistical products, users of these products depend on the attention given to curation. The effort of those who generate statistical products has a downstream impact not only on users but on the potential reuse and reputation of the shared statistical product. With the CDE focused on purpose-driven use, what needs to be considered by those who generate products to maximize their future use and impact and reduce the burden for future users?

The May 2022 Listening Session and stakeholder discussions yielded several challenges to explore for this topic:

10. Users, both internal and external to agencies, have difficulty finding and knowing how to access data products. How can the CDE and its curation processes be designed to ensure accurate (complete and up-to-date) inventories of data and statistical products available for reuse, thereby facilitating effective implementation of the “findable” and “accessible” principles?
11. Users depend on the documentation that describes a statistical product, how the product was generated, and for what purposes the statistical product is suited. It will be important to extend our knowledge of what data users require to understand, evaluate, and properly use shared data sources and documentation. Information scientists have investigated what users need to support the “reusable” principle and promote reproducibility (for example, see Faniel et al. 2010, 2016, 2017, and 2019, and Stodden 2015). How does existing knowledge impact the design of CDE products and documentation?
12. Users vary widely in their familiarity with and understanding of statistical programs that generate statistical products. In addition, their capacity to understand the technical underpinnings of the product range from nearly absent to being deeply familiar with technical

issues associated with the product. How can the wide variation in user skills and understanding of users be addressed (e.g., through user interfaces and types of products shared)?

13. More generally, the CDE should ideally engage with user communities to gain insight into the nature of users of statistical products, the quality and usability of statistical products, or the types of information needed by the public. What does this look like? National Academies of Sciences, Engineering, and Medicine (2021, p. 145, Recommendation 6.7) offers useful ideas for engaging with users. Another possibility is to consider the perspective of boyd and Sarathy (2022), who advocate for rethinking the purpose of the census and how data are governed.

Larger Trade-Offs

A final topic of exploration pertains to the costs and benefits associated with CDE curation. This is a difficult topic but is an important topic for the Census Bureau to address.

14. How can the trade-offs be described and studied?

Summary of Possible Research Areas

The previous section discussed ideas for incorporating curation into the CDE, as well as questions and challenges involved in developing a CDE curation approach. In this section, we summarize potential research areas, based on the questions and topics discussed above, that would facilitate developing effective and efficient curation processes for the CDE.

1. Explore state of the art application of FAIR principles in practice to identify their potential use in the CDE and map out an adoption strategy (Topic 5)
    - How are they appropriate to our application? When are they useful?
    - When are they challenging to apply?
    - What are successful examples of FAIR implementation, and why are they successful?
    - What does reusability mean in the CDE context?

2. Methods for promoting the potential for automation (Topic 2, 4)
    - What are the low hanging fruit for automating curation steps, and what will take more time?
    - What tasks have others been able to automate, and what does the Census Bureau already automate?
    - What factors make an element of the curation process amenable to automation, and can principles be developed to guide how we curate?
    - What is possible and valuable to harmonize?
    - How can AI be leveraged for lighter weight curation?
    - What kinds of tools or approaches can be used to develop documentation as a natural part of the production or ingestion process?

3. Identifying data quality assessment frameworks (Topic 6)
    - Explore data quality and other assessment frameworks
    - Develop a rubric/approach to evaluating different elements of data quality framework
    - Identify what should be documented and associated with a data source (or other object)

4. What are the models for the involvement of agencies, user communities, and data providers in the curation process? (Topic 7)
    a. Who has the right to do curation processes?
    b. How can studied communities be engaged in the curation process?
    c. What are appropriate roles for respondents in curation process, such as respondent-centered approaches in providing and withdrawing permissions for data to be used (for example, see NIH patient-centered outcomes research approaches)

5. Approaches to building partnerships with external sources (Topic 8)
    - What is the process for building the partnership?
    - What are the core elements of a value proposition for both entities?
    - What needs to be safeguarded and negotiated?
    - Identify successful partnership development activities as exemplars
    - Consider contract mechanisms, especially Other Transactions Authorities (OTAs) that support agile ways of responding to changing needs of project

6. Within the context of the CDE, what does it mean to curate? (Topic 1, 9, 11, 12)
    - Process – can this be broken down into phases of the process and what should be captured?
    - Purpose – what are the elements of purpose curation?
    - Fitness for purpose
    - Provenance – strategies for anticipating reuse, how to handle external sources with limited provenance information, how to handle dynamic or multiple data sources?
    - Linkages – from describing and standardizing to encryption
    - Terms of use, consent approaches in production and ingestion that enable future reuse
    - Dynamic data

7. How can the reuse of statistical products be facilitated so that it is appropriate and impactful? (Topic 11, 12, 13)
    - What documentation and context do users need to effectively reuse statistical products?
    - How can users with varying levels of knowledge, skills, and understanding be supported efficiently?
    - How can relationships with user communities be developed into an ongoing interaction?

8. How can the CDE curation framework be codified to support consistent and appropriate application of the framework? (Topic 1, 2)
    - How can the Census Bureau leverage publicly-available free/low-cost processing infrastructures (e.g., GitHub runners)?
        - Example applications: automated extraction of metadata on found sources, automating curation activities in Area 2
    - What tools are beneficial for tracking versions?

9. What standards should be adopted or developed for CDE curation? (Topic 3)
    - Investigate SDMX (Statistical Data and Metadata eXchange), which is used by international statistical agencies as a platform for cross-organizational harmonization
    - Unique resolvable persistent identifiers – what objects are tagged, and with what identifiers?

- What are practical approaches for developing metadata standards that will assist in automation and standardization? How to do this in an agile way?
  -

10. What is the infrastructure underpinning the CDE? (Topic 10, 11)
    - What are the access models for different levels of confidentiality (open, confidential non-Title 13, Title 13)?
    - What about platforms for serving metadata of digital objects:
    - How to reduce burden for users?

11. What are the economic implications of curation? (Topic 14)
    - How does CDE approach generate efficiencies and cost reductions?
    - Or are there new costs that need to be planned? Who pays for extra costs?

12. Census Bureau inventories and evaluations (Topic 10)
    - Computing infrastructure to support volume and variety of data that could be accessed and implementation of curation activities
    - Census Bureau-wide administrative data processes and experiences
    - Partnership development to obtain external data
    - Internal processes for ingesting and documenting external or internal data

Summary

This white paper outlines a perspective on curation for the Curated Data Enterprise that focuses on transparency and agile reuse of data sources for the public good. We developed this perspective based on the listening sessions with stakeholders, discussions with Census Bureau staff, and information from scholarly literature.

We introduce the CDE curation framework for considering a deeper and more complete form of curation than is typically practiced. This framework has been tested extensively at the Social and Decision Analytics Division of the UVA Biocomplexity Institute through their experience curating many different sources of public and private data for a vast range of purposes that benefit the public.

To provide some specificity on a form of CDE curation that prioritizes transparency and reuse, we propose a definition for curation that explicitly focuses on the transparency of purpose and use, processes, and products. We also outline specific types of documentation that foster transparency and reduce effort in reuse of statistical products. Our definition of curation calls for documenting each step of the CDE framework. This includes the purpose, inputs (e.g., data, evaluations, decisions), and manipulations (e.g., software) used to generate a statistical product, which may be data, visualizations, reports, or other digital objects. Many of these curation steps are illustrated in [cite SNF paper].

Our discussions with stakeholders exposed several areas for exploration in developing a curation process for the CDE. We grouped these topics to generate a roadmap of 12 potential areas for future research that would be beneficial in developing effective and efficient curation processes for the CDE.

Further work is needed to prioritize these areas to create research projects that would provide substantive insights in short order to continue refining the CDE curation concept. An agile research process that involves a series of smaller steps will provide an efficient way to further specify the CDE curation approach in a way that focuses on purpose-driven creation and reuse of statistical products.

References

boyd d, Jayshree, S. 2022. Differential perspectives: epistemic disconnects surrounding the US Census Bureau's use of differential privacy. *Harvard Data Science Review*, (Special Issue 2). https://doi.org/10.1162/99608f92.66882f0e

Dempsey, WP, Foster I, Fraser S, Kesselman C. 2022. Sharing begins at home: how continuous and ubiquitous FAIRness can enhance research productivity and data reuse. *Harvard Data Science Review*, 4(3). https://doi.org/10.1162/99608f92.44d21b86

Faniel IM, Frank RD, Yakel E. 2019. Context form the data resuer's point of view. *Journal of Documentation*, 75(6):1274-1297. https://doi.org/10.1108/JD-08-2018-0133

Faniel IM, Jacobsen TE. 2010. Reusing scientific data: how earthquake engineering researchers assess the reusability of colleagues' data. *Computer Supported Cooperative Work* 19:355-375.

Faniel IM, Kriesberg A, Yakel E. 2016. Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology.* https://doi.org/10.1002/asi.23480

Faniel IM, Yakel E. 2017. Practices do not make perfect: disciplinary data sharing and reuse practices and their implications for repository data curation. In: *Curating Research Data, Volume 1: Practical Strategies for your Digital Repository.* Johnson LL, Ed. Chicago: Association of College and Research Libraries.

Keller, S., Prewitt, K., Thompson, J., Jost, S., Barrett, C., Nusser, S., Salvo, J., Shipp, S. (2022). A 21st Century Census Curated Data Enterprise. A Bold New Approach to Create Official Statistics. Proceedings of the Biocomplexity Institute. Technical Report 2022-1115. https://doi.org/10.18130/r174-yk24

Mikytuck AM, Nusser SM, Kormaz G. 2022. The interdependence of data producers and data users: How researchers' behaviors can support or hinder each other. Submitted to *PLOS ONE*. Preprint available at https://osf.io/preprints/metaarxiv/yp3ct/ (last updated on October 24, 2022).

National Academies of Sciences, Engineering, and Medicine. 2021. *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and all Federal Statistical Agencies.* Washington, DC: The National Academies Press. https://doi.org/10.17226/26360

National Academies of Sciences, Engineering and Medicine. 2022. *Automated Research Workflows for Accelerated Discovery*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26532

Nusser SM. 2023. The role of statistics in promoting data reusability and research transparency. *Annual Review of Statistics and Its Applications*, 10:1-23. https://doi.org/10.1146/annurev-statistics-033121-105114

Poole AH. 2016. The conceptual landscape of digital curation. *Journal of Documentation*, 72(5):961-986. https://doi.org/10.1108/JD-10-2015-0123

Shipp S, Zhu Z, Jost S, Naymark J, Prewitt K, Salvo J. Thompson J, Snipp M, Becker-Medina E. (2024). Developing a 21st Century Census Curated Data Enterprise: A Bold New Scientific Approach for Official Statistics. University of Virginia. DOI: https://doi.org/10.18130/sg02-cy43

Stodden, V. 2015. Reproducing statistical results. *Ann. Rev. Stat. Appl.* 2:1-19.

Willis C, Stodden V. 2020. Trust but verify: how to leverage policies, workflows, and infrastructure to ensure computational reproducibility in publication. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.25982dcf