TOOL 5

Excerpted from Wiegand, E.R., Monahan, E.K., Geoghegan, R., Wavelet, M., & Goerge, R.M. (2023.) *Strengthening Analytics in Government Agencies: A Toolkit for Sustainable Data Use*. OPRE Report 2023-148. Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

A DOCUMENTATION CHECKLIST

One of the largest challenges agencies face in organizing and documenting information is thinking about everything that needs to be documented to ensure sustainable data use and the preservation of institutional memory. An organization might have great systems for documenting processes, projects, or data, for example, and still find itself lacking critical information in other areas. In fact, strong documentation in one area can even lead a team to become complacent and overlook other important areas.

This tool details questions that highlight the breadth of knowledge that may be important for a state agency analytics team to track. It was developed based on the research team's experience working with state agencies to build analytic capacity and later revised based on feedback from the toolkit interviewees. Depending on the size of your team and the scope of its work, you may choose to prioritize the quality of your documentation in some subset of these areas: Documentation of Source Data Systems; Documentation of Commonly Used Data Extracts, Views, or Datasets; Documentation of Frequently Used Definitions; Contextual Information About Data Collection, Quality, and Policy; Documentation of Analytics Workflows, Decisions, and Products. This tool is intended to highlight those choices for your consideration.

INSTRUCTIONS

Review the questions below and think about the tools and processes used to document this information at your agency. In each area, consider:

- Is there any documentation on this subject in your agency?
- Is there an agreed upon "single source of truth" for this information? Are there processes and training in place so users routinely update the single source of truth when they learn something new or when something changes?
- Is the documentation searchable and interpretable by most users? Do users know about the documentation and consistently access it when they should?

Once you have reviewed the full list of documentation needs, think about what your organization's next documentation priorities should be. **Balancing the importance of the information against the quality of current documentation, which areas are priorities for improved documentation processes or tools?**

Documentation of Source Data Systems

- Which of the following are available? When and how are they updated?
  - Table and field lists
  - Entity relationship diagrams
  - Data dictionaries (including field names, definitions, variable types, and indicators for primary keys)
  - Codebooks

Documentation of Commonly Used Data Extracts, Views, or Datasets

- Are there data dictionaries or codebooks for these resources?
- Where is the source code or logic used to create the data stored? Is the source code well-commented and easily understandable to at least a technical audience?
- How are dataset versions handled? Ideally, there is clear documentation of dataset versions, along with associated code and dates run (for example, if a dataset changes, is it easy to determine who made the change, when, and why?)

Documentation of Frequently Used Definitions

- How are key concepts (for example, case types, exits, program participation status, total earnings) defined in the data?
  - Are there authoritative code snippets or sets of logic to define key concepts? How are these maintained and tracked over time?
  - If the logic to define these concepts varies, is there clear documentation of where and why?

Contextual Information About Data Collection, Quality, and Policy

- What documentation exists showing how data are collected, and what guidance is given to staff members doing data collection or data entry about how different situations are coded? Relevant information might be captured in user manuals and training materials.

- As analysts or policy staff members identify problems in data elements (such as missingness, inaccuracies in data collection, or variation over time), are these findings documented?

- Changes in areas such as program eligibility or requirements impact what patterns are expected in the data. Is there consistent documentation of policy changes and their implications for the data?

Documentation of Analytics Workflows, Decisions, and Products

- Is there documentation of which analyses have been performed, either routinely or on an ad hoc basis? Can staff members easily build on what has been done before, or are they constantly called on to recreate logic, code, or analyses?

- Is there documentation of key workflows (for example, how are scripts, datasets, and result files related)? For a routine analysis, if the person who runs the report is out unexpectedly, could someone else pick it up? For an ad hoc analysis, could someone replicate the analysis months or years in the future? Process maps and clear naming conventions can help make workflow components more accessible.

- Are there log files or other records of when (and by whom) files were last changed or run? Version control tools such as Git can help track changes in code. Are final versions clearly delineated?

- Are there records of key analytics decisions (for example, why it was done the way it was done)?

- Is the original data behind any analysis clearly documented, including both the extract date and the source system?