SOCIAL WELLBEING AGENCY | TOI HAU TĀNGATA

Information sheet

Accelerating dataset assembly

December 2020

New tool and process for dataset preparation

Data preparation for analytic and research projects increases in complexity as the number of data sources increases. Without a consistent method, it can become time-consuming, expensive, and error prone.

The Social Wellbeing Agency has developed the Dataset Assembly Tool. By standardising and automating dataset assembly, the tool helps staff deliver higher quality work faster.

The Dataset Assembly Tool is now available for other researchers and analysts to use.

It is supported by extensive documentation, worked examples, a range of optional patterns, and a library of existing definitions. With these resources, new staff have become comfortable using the tool within two days with minimal training.

The value of faster dataset preparation

The Social Wellbeing Agency has already found the Dataset Assembly Tool to be valuable:

- **Accelerated assembly** – The assembly tool more than halves the time to create research ready datasets, considerably reducing the cost to commence research. In one project SWA staff produced a panel dataset from 25 sources within a week of technical work beginning.
- **Rapid iteration** – A faster process supports ongoing improvement. In another project SWA staff updated the primary dataset every hour as staff collaborated checking and polishing the dataset.

The value of the tool is not limited to these examples:

- **Fewer errors** – A standardised process can be rerun with confidence that it will continue to perform. It also reduces the number of places where errors can occur, reducing the time spent finding faults.
- **Scalable development** – By encouraging independence between inputs, the tool enables researchers to work in steps that are easy to understand and manage.

- **Single step construction** – As the research dataset is combined together in one step, staff avoid the hazard of tangled, bespoke assembly patterns. The tool can be a starting point for data workflows.
- **Definition reuse** – The measures defined for a project represent expert knowledge and have value beyond the project they were created for. By separating definitions from assembly, the tool enables easy reuse of definitions between projects and sharing of definitions with other researchers.
- **Increased collaboration** – By providing standard patterns, the tool encourages consistency between staff, making collaboration on a single project, or handover of projects more straightforward.
- **New opportunities for innovation** – When consistent processes are adopted, opportunities arise for further improvement across staff and projects by spreading good practices and resources.

Further resources are available

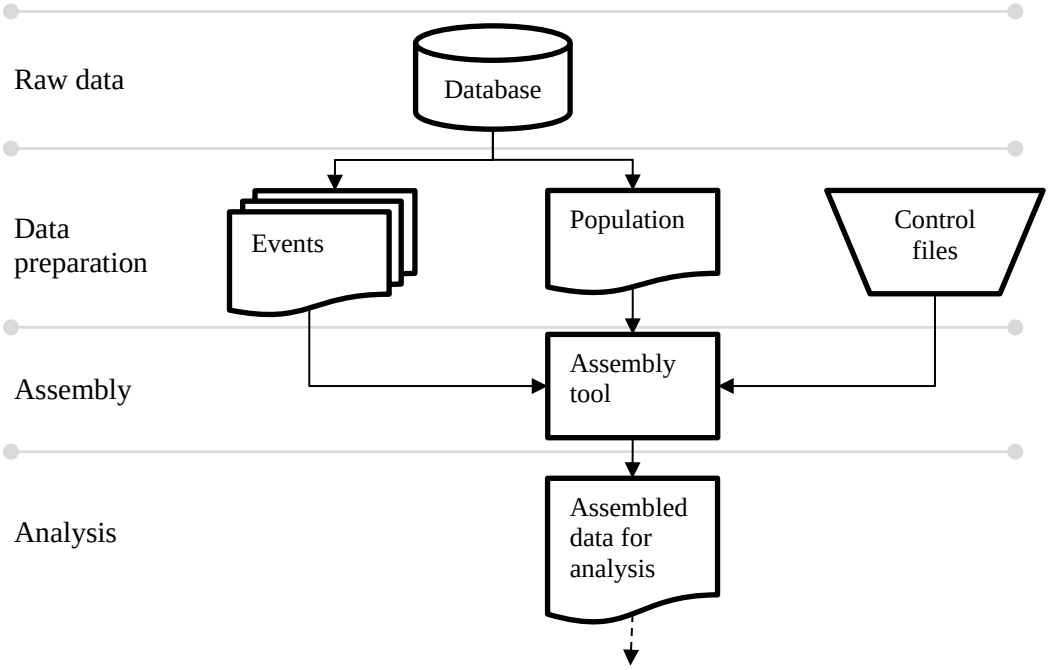
Interested staff are encouraged to review the supporting documentation:

The primer and guide to the assembly tool. This covers an introduction to the tool, recommends an effective project structure and workflow, gives a description of the design, and concludes with a worked example.

The training presentation. Available as a video, this introduces the tool, describes how to install it, provides an example of configuring and running the tool, and gives guidance on fixing errors that might arise during use. The presentation contains extensive speakers notes that serve as a users' manual.

The GitHub repository for the tool. This contains all the code required to run the tool, automated tests for validating performance, and examples of using the tool.

Figure: Flow of information when using the Dataset Assembly Tool



https://github.com/nz-social-wellbeing-agency/dataset_assembly_tool