What is Data Curation and Preservation?

Data curation and preservation is the art of maintaining the value of data. A data curator does this by collecting data from many different sources and then aggregating and integrating it into an information source that is many times more valuable than its independent parts. During this process, data might be annotated, tagged, presented, and published for various purposes. The goal is to keep the data valuable so it can be reused in as many applications as possible.

Through the curation process, data are organized, described, cleaned, enhanced, and preserved for public use, much like the work done on paintings or rare books to make the works accessible to the public now and in the future. With modern technology, it's increasingly easy to post and share data. Without curation, however, data can be difficult to find, use, and interpret.

Producers of official data (such as national statistical agencies, line ministries, development projects) generate vast amounts of digital information each year, including microdata from surveys, analytical reports, methodological documents and questionnaires. These digital assets represent significant investment by producers and therefore need care and attention to survive. They need to be preserved to ensure that they also benefit future researchers. Unfortunately, there are numerous cases where important data has been lost in the absence of an effective preservation approach. For example, data have been stored on outdated, unreadable mainframe tapes and other obsolete technology.

When a dataset is no longer being used, it is most at risk and needs long-term preservation. This process requires introduction of related standards and good practices. Data preservation is the act of conserving and maintaining both the safety and integrity of data. Preservation is done through formal activities that are governed by policies, regulations and strategies directed towards protecting and prolonging the existence and authenticity of data and its metadata. Data can be described as the elements or units in which knowledge and information is created, and metadata are the summarizing subsets of the elements of data; or the data about the data. The main goal of data preservation is to protect data from being lost or destroyed and to contribute to the reuse and progression of the data.

Data preservation consists of a series of managed activities necessary to ensure continued access to data for as long as necessary. It is not just primarily about conservation or restoration, storage media or backup regimes or concepts of “permanence”. Data preservation requires ongoing active management of data from as early in the lifecycle as possible and is as much a management function as it is a technical activity.

As the volumes, and complexity of data grows, active management of data preservation becomes a very important component. Staff in National Statistical Offices (NSO) who work closely with data and the research community can play a major role in the ongoing active management and preservation of data by ensuring that microdata is arranged and described with appropriate metadata, stored in a location where it can be monitored and made available for future use, and prepared for migration or transformation if the data format is damaged or in an obsolete format.

Why is digital data preservation important?

Unlike the preservation of information on paper, the preservation of digital information demands constant attention. Long-term storage of development data demands additional efforts and costs for their preparation in the format that will enable further use. These costs and efforts are justified by the savings, represented in the continuous reuse of data.

Microdata preservation also includes management of related metadata over time to guarantee their long-term usability. Organizations which disseminate microdata and the related metadata are also often responsible for their preservation. Preserving digital content is not a trivial exercise. It requires the establishment and implementation of a preservation policy and procedures to ensure that data and all related metadata are preserved against:

- hardware or software obsolescence,
- media failure, and
- other physical threats.

Procedures and infrastructures must be put in place to protect data against hardware and software obsolescence (regular migration of datasets to new media and formats), system failures, human errors and other hazards. Micro-datasets can be damaged or lost because of human error, because of technical problems that lead, for example, to the corruption of data files, or because of disasters such as earthquake, fire or flood. New technologies can also render old data unreadable, because of either hardware or software advances. The importance of preserving data is vast. When data is lost it is as though it never existed. It is important to realize that data is the building block of everything, it is seen on both small and large scales.

For example data often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects may analyse or add to the data, and data may be re-used by other researchers. Well organised, well documented, preserved and shared data are invaluable to advance scientific inquiry and to increase opportunities for learning and innovation.

The Pacific Community (SPC) Pacific Data Hub Microdata Library's curation team preserves data by utilizing tools like Stat Transfer, Nesstar Publisher and Stata to convert microdata files to usable formats (earlier versions of Stata, SPSS, CSV, ASCII, SAS and SPSS syntax), by providing variable and value labels, and archiving the data in a safe backed-up, storage space on the network drive.