Scientific data chain for better data publication and preservation

*Takeshi Horinouchi[1]

1. Faculty of Environmental Earth Science, Hokkaido University

In this presentation, I will discuss some of the current issues and thoughts on data publication and preservation from the viewpoint of an atmospheric scientist in Japan. Earth observation data are not only experimental data but also are historical data, so their preservation and sharing are no less important than those from laboratory experiments. However, the current status is not satisfactory. I will introduce a way that might improve the situation.

In our field, we have “big” data and “small” data in two aspects. Data from big projects such as satellite observations, coordinated observations for weather forecast, climate prediction for IPCC are big in size and well preserved and shared. On the other hand, data from small projects such as field observations are not necessarily well-shared, and their preservation may be at risk. These longtail data are also important. For them, use of public or institutional data repository (and data journal) may be a solution, but to use it (them) is not yet our common conduct. More incentive and familiarity would be needed.

The other aspect of the big-and-small issue is on the level of process. For example, numerical values visualized in figures (such as line charts) in peer-reviewed papers can be regarded to be on a highest process level. They are normally small and derived from lower-level data, which can be big. This process is usually multi-staged, and to reproduce it is often difficult even when the original data are open.

Considering these issues, one can envision a digital scientific data chain (like food chain) in which research paper is on the top. Here, a plot in a figure is linked to a digital table (like CSV) stored in journal’s site. The table is further linked to a data repository content or a data paper (through which data repository can be linked). A well-organized data chain can realize reproducibility through documentation and digital operations. The latter can be helped by scripting. An example would be to script data screening and processing such as averaging. The data link in this chain can be multifold, starting from big low-level data.

This chain will make scientists familiar with data repository and data paper. It can also give incentive to write data paper to cite, so it can encourage documentation that increases the usability of data. It can also provide a good example of data usage. Therefore, it can promote data publication and preservation.

Keywords: data publication, open data

