**Data lifecycle is not a cycle, but a plane!**

***Javad Chamanara1, Birgitta König-Ries1***

**1***Friedrich Schiller University Jena, Germany*

Corresponding author(s) e-mail: javad.chamanara@uni-jena.de

**ABSTRACT:** Most of the data-intensive scientific domains, e.g., life-, natural-, and geo-sciences have come up with data life cycles. These cycles feature, in various ways, a set of core data-centric steps, e.g., planning, collecting, describing, integrating, analyzing, and publishing. Although they differ in the steps they identify and the execution order, they collectively suffer from a collection of short-comings.

They mainly promote a waterfall-like model of sequentially executing the lifecycles’ steps. For example, the lifecycle used by DataOne suggests that “analyze” happens after "integrate". However, in practice, a scientist may need to analyze data without performing the integration. In general, scientists may not need to accomplish all the steps. Also, in many cases, they simply jump from, e.g., "collect" to "analyze" in order to evaluate the feasibility and fitness of the data and then return to "describe" and "preserve" steps. This causes the cycle to gradually turn into a mesh. Indeed, this problem has been recognized and dealt with by the GFBio and USGS data lifecycles. The former has added a set of direct links between non-neighboring steps to allow shortcuts, while the later has factored out cross-cutting steps, e.g., "describe" and "manage quality" and argued that these tasks must be performed continually across all stages of the lifecycle. Although aforementioned lifecycles have realized these issues, they do not offer customization guidelines based on, e.g., project requirements, resources availability, priority, or effort estimations.

In this work, we propose a two-dimensional Cartesian-like plane, in that the x- and y-axes represent phases and disciplines, respectively. A phase is a stage of the project with a predefined focus that that leads the work towards achieving a set of targeted objectives in a specific timespan. We identify four phases; conception, implementation, publishing, and preservation. Phases can be repeated in a run, and do not need to have equal timespan. However, each phase should satisfy its exit criteria to be able to proceed to the next phase. A discipline, on the vertical axis, is a set of correlated activities that, when performed, makes a measurable progress in the data-centric project. We have incorporated these disciplines: plan, acquire, assure, describe, preserve, discover, integrate, analyze, maintain, and execute.

An execution plan is developed by placing required activities in their respective disciplines’ lanes on the plane. Each task (activity instance) is visualized as a rectangle that its width and height respectively indicate the duration and effort estimation needed to complete it. The phases, as well as the characteristics of the project (requirements, size, team, time, and budget), may influence these dimensions.

It is possible for a discipline or an activity to be utilized several times in different phases. For example, a planning activity gains more weight in conception and fades out over the course of the project, while analysis activities start in mid-conception, get full focus on implementation, and may still need some attention during publishing phases. Also, multiple activities of different disciplines can run in parallel. However, each task's objective should remain aligned according to the phase’s focus and exit criteria. For instance, an analysis task in the conception phase may utilize multiple methodologies to perform experimentation on a small sample of a designated dataset, while the same task in the implementation phase conducts a full-fledged analysis using the chosen methodology on the whole dataset.

**KEYWORDS:** Data Lifecycle, Data Management, Research Data Management, Scientific Data Management

**REFERENCES: (ONLY if applicable)**