A cost-based ratio scale for data FAIRness

Yoram Timmerman[1], Antoon Bronselaer[1], Filip Pattyn[2]

[1]) Department of Telecommunications and Information Processing, Ghent University, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
yoram.timmerman@ugent.be, antoon.bronselaer@ugent.be
[2]) Ontoforce, Technologiepark 122, 9052 Gent, Belgium
filip.pattyn@ontoforce.com

The recent dramatic increase in the globally created amount of data has raised numerous questions regarding the quality of these data and, more specifically, how to measure this quality. Although quality measurement procedures do exist, we argue that they typically come with several drawbacks. First, data quality is typically measured on an ordinal scale. Such a quality scale is typically constructed based on a couple of criteria to which good quality data should adhere. However, an ordinal scale does not allow to deduce *how much* better a data set is compared to another. Moreover, the larger the list of criteria becomes, the more difficult it becomes to construct a reliable measurement system. Finally, the criteria that are typically used to assess data quality are purely intrinsic properties of the data set itself. However, one could argue that it is not possible to assess the perceived quality of a data set without looking at the context in which the data are used.

In the first part of this talk, a new way of looking at data quality will be discussed. We argue that it is important to take into account the specific purposes for which the data are used when measuring their quality. As such, in our approach data quality is evaluated by looking at the amount of effort that is needed to complete a predefined set of *tasks* for which the data are needed. This effort will be quantified in terms of a *cost*, leading to an intuitive ratio scale for data quality instead of an ordinal scale.

In the second part of this talk, a practical illustration of the idea of cost-based data quality measurement will be given. More specifically, it will be illustrated how cost-based measurement can be used to compare the FAIRness of data sets. The FAIR Data Principles were initially designed as guidelines to stimulate proper scientific data management and to promote data reuse [1]. However, no consensus has been found yet on how the FAIRness of a data set should be measured. Constructing an ordinal quality scale based on the different FAIR criteria would probably be very difficult because of the long checklist of criteria that are part of the FAIRness definition. Contrary, cost-based quality measurement would allow to express how much effort is needed to make a data set FAIR enough to be able to perform a set of tasks. The idea behind cost-based FAIRness measurement will be made clear with the help of a real-life biomedical use case that should be solved using two different data sets.

References

[1] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3 (2016).