# Whip: Human and machine-readable specifications for data

***Stijn Van Hoey*[1], *Peter Desmet*[1]**

[1]*Research Institute for Nature and Forest (INBO), Belgium*

Corresponding author(s) e-mail: stijn.vanhoey@inbo.be

**ABSTRACT:** Different tools and technologies are available to clean and harmonize data. Independent of the tool used, the ability to assess the quality of a data set and identify potential errors is crucial for harmonization efforts. The necessity becomes even more apparent in the context of data publication, (re)use and aggregation.

Documentation and guidelines about the data requirements provide guidance in this process and enable to communicate what to expect from the data, but are mostly intended for humans only. To facilitate the harmonization process, we propose the usage of a specification file, describing the constraints to which the data should comply. Its syntax is human- and machine-readable, so it can be used to communicate expected data quality/conformity and to validate data automatically. The scope of the set of specifications can be specific to a dataset, researcher or research community, which allows bottom-up and top-down adoption. As an example, we apply the specifications to verify data mapped to the biodiversity information standard Darwin Core.

In this talk, we will present "whip", a proposed syntax and format to express data specifications. Whip allows to define column-based constraints for tabular (tidy) data with a number of rules. We will also demonstrate a software application (called "pywhip") to validate data sets using these specifications. We hope it will trigger a discussion on how to express data specifications and communicate data quality expectations.

**KEYWORDS:** Data Harmonization, Data Quality, Documentation, Specifications