Geophysical Research Abstracts
Vol. 18, EGU2016-**PREVIEW**, 2016
EGU General Assembly 2016

An algebra for spatio-temporal information generation

Edzer Pebesma (1), Simon Scheider (2), Benedikt Gräler (3), Christoph Stasch (4), and Matthias Hinz (1)

(1) Institute for Geoinformatics, University of Münster, Germany (edzer.pebesma@uni-muenster.de), (2) Human Geography and Spatial Planning, Utrecht University, The Netherlands, (3) Institute of Hydrology, Ruhr University Bochum, Germany, (4) 52North GmbH, Münster, Germany

When we accept the premises of James Frew's laws of metadata (Frew's first law: scientists don't write metadata; Frew's second law: any scientist can be forced to write bad metadata), but also assume that scientists try to maximise the impact of their research findings, can we develop our information infrastructures such that useful metadata is generated automatically? Currently, sharing of data and software to completely reproduce research findings is becoming standard, e.g. in the Journal of Statistical Software [1]. The reproduction (e.g. R) scripts however convey correct syntax, but still limited semantics.

We propose [2] a new, platform-neutral way to algebraically describe how data is generated, e.g. by observation, and how data is derived, e.g. by processing observations. It starts with forming functions composed of four reference system types (space, time, quality, entity), which express for instance continuity of objects over time, and continuity of fields over space and time. Data, which is discrete by definition, is generated by evaluating such functions at discrete space and time instances, or by evaluating a convolution (aggregation) over them. Derived data is obtained by inputting data to data derivation functions, which for instance interpolate, estimate, aggregate, or convert fields into objects and vice versa.

As opposed to the traditional *when, where* and *what* semantics of data sets, our algebra focuses on describing *how* a data set was generated. We argue that it can be used to discover data sets that were derived from a particular source x, or derived by a particular procedure y. It may also form the basis for inferring meaningfulness of derivation procedures [3]. Current research focuses on automatically generating provenance documentation from R scripts.

[1] http://www.jstatsoft.org/ (open access)

[2] http://www.meaningfulspatialstatistics.org has the full paper (in review)

[3] Stasch, C., S. Scheider, E. Pebesma, W. Kuhn, 2014. Meaningful Spatial Prediction and Aggregation. Environmental Modelling & Software, 51, 149–165 (open access)