

Data Quality and Fitness for Purpose

Franz-Benjamin Mocnik, Hongchao Fan, Alexander Zipf
GIScience Research Group, Institute of Geography, Heidelberg University

This poster has been supported by the DFG project *A framework for measuring the fitness for purpose of OpenStreetMap data based on intrinsic quality indicators* (FA 1189/3-1).



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

... DATA QUALITY AND FITNESS FOR PURPOSE
STAY OFTEN UNRELATED ...

FITNESS FOR PURPOSE

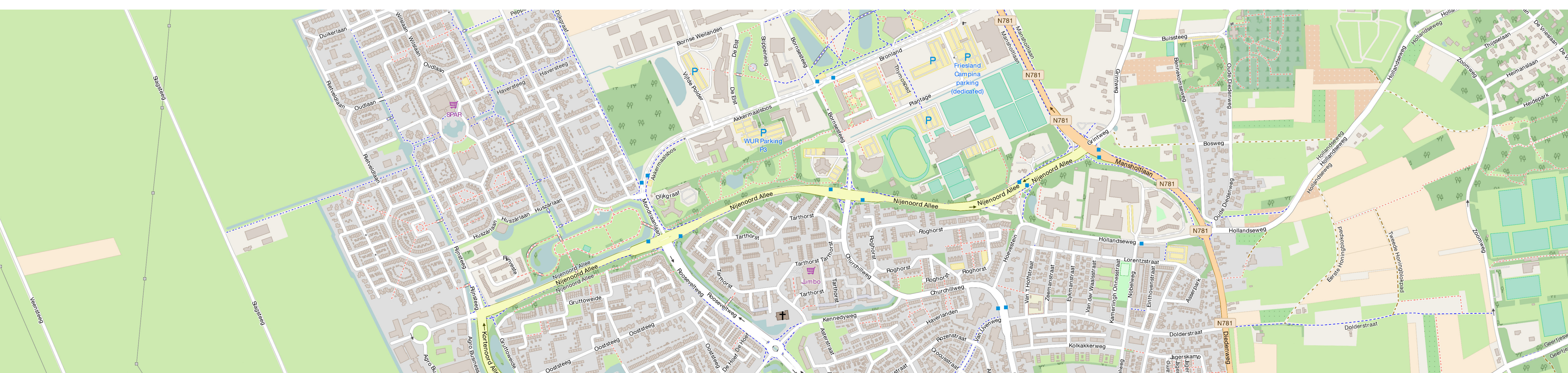
Fitness for purpose is the affordance of data to be interpreted and used in a context that renders a certain usage, that is, the purpose, possible.

DATA QUALITY

Data quality is the discrepancy between the fitness for purpose of optimal data, that is, of data with maximal fitness for purpose but the same scope as the actual data, and the fitness for purpose of the actual data, aggregated for all possible purposes.

The map below is not fit for the purpose of deriving a soil classification (because soil types are not depicted) ...

... but might still be of good quality!



Data Quality and Fitness for Purpose

Franz-Benjamin Mocnik
Heidelberg University
GIScience Group
Im Neuenheimer Feld 348
69120 Heidelberg, Germany
mocnik@uni-heidelberg.de

Alexander Zipf
Heidelberg University
GIScience Group
Im Neuenheimer Feld 368
69120 Heidelberg, Germany
zipf@uni-heidelberg.de

Hongchao Fan
Heidelberg University
GIScience Group
Im Neuenheimer Feld 348
69120 Heidelberg, Germany
hongchao.fan@uni-heidelberg.de

Abstract

Different definitions of data quality and fitness for purpose exist in literature. We propose definitions that align well with existing definitions but emphasize how both concepts relate.

Keywords: Data quality; fitness for purpose

1 Introduction

In literature, the definitions of fitness for purpose and data quality come in two guises: data quality defined by internal characteristics, for example, by completeness, logical consistency, positional and thematic accuracy, temporal quality, etc. (International Organization for Standardization); and data quality defined in terms of the use of the data (Chrisman, 1984; Frank, 1998; Frank, 2009). While there seems to be an agreement on the fact that concepts entitled ‘data quality’ and ‘fitness for purpose’ exist, no clear distinction between them is made in many cases. Barron et al. (2014) state, for example, that

‘OSM data quality heavily depends on the purpose for which the data will be deployed. We refer to this as “Fitness for Purpose” assessment, previously defined by Veregin [...] as determining “fitness-for-use”.’

In a similar way, Chrisman (1984) states that

‘Quality information provides the basis to assess the fitness of the spatial data to a given purpose’.

Both statements demonstrate that data quality and fitness for purpose are commonly understood as being closely related, yet an understanding of how they formally relate is missing. In fact, data quality and fitness for purpose shed light on the same problem from different points of views (Devilleers, et al., 2005).

The poster aims at proposing and discussing definitions for the terms ‘data quality’ and ‘fitness for purpose’, such that

- (1) both terms are defined in common terms, that is, referring to a common vocabulary, making both definitions ‘compatible’,
- (2) the definitions relate both terms,
- (3) the definitions align well with existing definitions and common usages of the terms, at least in realistic scenarios, and
- (4) such that the definitions are independent of a particular data model.

2 Definitions

Data get a meaning if they are interpreted: symbols like characters, words, or numbers, become related to the environment. Such interpreted data have become information, because we have an understanding of how to make use of the data for tackling problems, for example, for solving route planning tasks.

Different interpretations of the same data can lead to different information. The interpretation of a map as a cyclist or as a driver of a motor vehicle may, for example, lead to different information about distances. When data is interpreted in a suitable context, the gained information may be used to solve a given task, and the potential to solve this task depends on both, the data and its interpretation. Route planning tasks can, for example, only be performed, if the map is suitable for this purpose and if the reader knows how to interpret the map in a suitable context. It can be seen as a feature of the data, called an ‘affordance’, that the data can be interpreted in a suitable context, acting as an ‘environment’ of the data (Gibson, 1977; Sanders, 1997; Turvey, 1992). We can define the fitness for purpose, that is, how fit data is to be used for a certain purpose, in terms of affordances: *fitness for purpose is the affordance of data to be interpreted and used in a context that renders a certain usage, that is, the purpose, possible.*

The fitness for purpose of data refers to a certain purpose of how to use the data, but data is often to be assessed without having any particular purpose in mind, for example, to measure the quality of the data: How complete and how consistent are the data? How precise are the coordinates or the semantic information? etc. Data quality does not, in contrast to fitness for purpose, measure how well-suited data is for a certain purpose, but whether it meets our *expectations* when being used for *different purposes*. A map or dataset may, for example, be of high quality, despite not representing soil types and thus not being fit for the purpose of deriving a soil classification. Data quality is independent of a particular purpose, because it is assessed in respect to all possible purposes. We accordingly define: *data*

quality is the discrepancy between the fitness for purpose of optimal data, that is, of data with maximal fitness for purpose but the same scope as the actual data, and the fitness for purpose of the actual data, aggregated for all possible purposes. High data quality indicates a small discrepancy, and low data quality, a large one.

Acknowledgement

This work has been partially supported by the DFG project *A framework for measuring the fitness for purpose of OpenStreetMap data based on intrinsic quality indicators* (FA 1189/3-1).

References

- Christopher Barron, Pascal Neis, and Alexander Zipf. A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, 18(6):877–895, 2014.
- Nicholas R. Chrisman. The role of quality information in the long-term functioning of a geographic information system. *Cartographica*, 21(2):79–87, 1984.
- Rudolphe Devillers, Yvan Bédard, and Robert Jeansoulin. Multidimensional management of geospatial data quality information for its dynamic use within GIS. *Photogrammetric Engineering and Remote Sensing*, 71(2):205–215, 2005.
- Andrew U. Frank. Metamodels for data quality description. In Robert Jeansoulin and Michael F. Goodchild, editors, *Data quality in geographic information. From error to uncertainty*, page 15–29. Hermès, Paris, 1998.
- Andrew U. Frank. Why is scale an effective descriptor for data quality? The physical and ontological rationale for imprecision and level of detail. In Gerhard Navratil, editor, *Research trends in geographic information science*, page 39–61. Springer, Heidelberg, 2009.
- James J. Gibson. The theory of affordances. In Robert Shaw and John Bransford, editors, *Perceiving, acting, and knowing*, page 127–143. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.
- International Organization for Standardization. ISO 19157:2013: Geographic information. Data quality.
- John T. Sanders. An ontology of affordances. *Ecological Psychology*, 9(1):97–112, 1997.
- Michael T. Turvey. Affordance and prospective control: An outline of the ontology. *Ecological Psychology*, 4(3):173–187, 1992.