Quality Assessment of Big Data with GIS

Michiel Jellema
Infofolio BV

P.O. Box 72
3700 AB Zeist
The Netherlands
m.jellema@infofolio.nl

Marinus de Bakker
HAS University of
Applied Science
P.O. Box 90108
5200 MA Den Bosch
The Netherlands
m.debakker@has.nl

Abstract

The increase of available data and more users with different needs influence the approach regarding data quality. We propose an integration of Big Data with data quality assessment. This Big Data Quality Assessment Model (BDQAM) is illustrated by a case of Infofolio where GIS functionalities play an important role. We conclude that although the approach is promising, more research is needed regarding the data quality dimensions of Big Data and the relationship between GIS and Big Data.

Keywords: Information Society, Data Quality Assessment, Big Data, GIS, Fitness for use, Data Reference Quality

1 Introduction

The information society develops rapidly. More data are collected via network connected sensors. This development, called 'Big Data' refers not only interpretation of large data sets, but also smart combining and fusion of different data sources [Klous et al, 2016].

The 'fitness for use' principle will be more difficult to assess and assure if many different (big) data sources are used. So not only the original owner of the data is responsible, but also all the stakeholders involved in upgrading the data towards an objective level of data quality for the end-users. This objective level of data quality is defined as the data reference quality. The data reference quality describes the optimal data quality of every data attribute, considering the possibilities of the original data sources.

Based on the theoretical background of the terms 'data quality' [Cai et al, 2015], [DAMA, 2013], 'big data' [Katal, 2013] and 'GIS' [By de, R.A. et al., 2001], the characteristics and possibilities of the model 'Big Data Quality Assessment Model' [Jellema, De Bakker, 2017] are introduced.
This model will be applied to an example of the daily data process of Infofolio [Infofolio, 2017]. Making use of the possibilities of Big Data and multiple linear regression analysis [Jellema et al, 2015], the data from over more than 50 different data sources are collected, analysed and made usable by Infofolio, in more than 150 data attributes on address-level of almost 9 million buildings, for organisations in the insurance-, risk- and safety-branches.

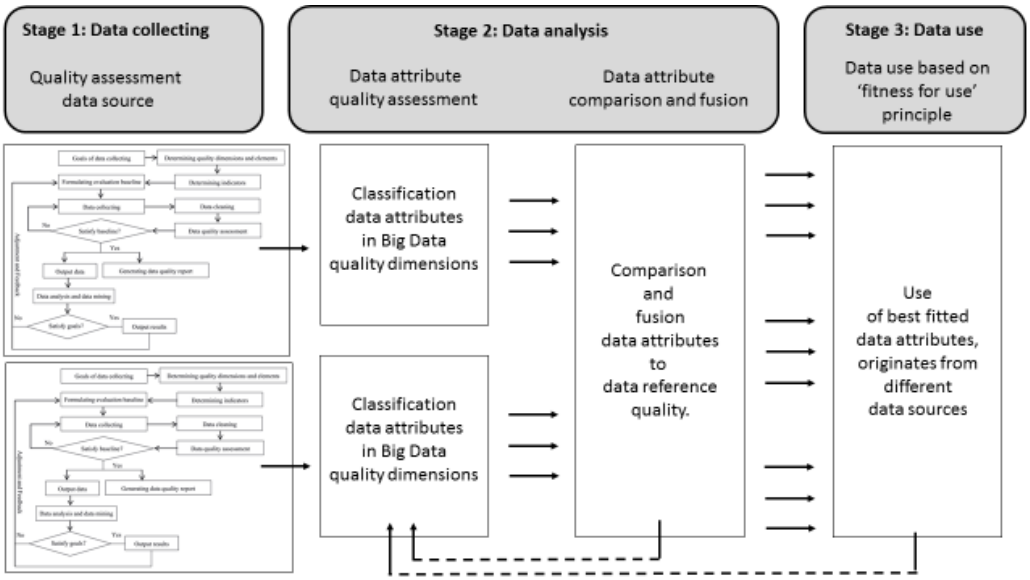
2 Big Data Quality Assessment Model

This 'Big Data Quality Assessment Model' (figure 1) is a further development of the model of Cai and Zhu (2015) with the following aspects:

- The model divided the Big Data processes in a three-stage procedure: The process from Cai and Zhu (2015) takes place in the first stage procedure. The quality assessment process is executed for every individual dataset.
- In the second stage procedure, the outcome of every dataset is classified per data attribute in the Big Data quality dimensions and comparison is possible per comparable data attribute of the different datasets. After the quality assessment on the level of data attributes, the outcome will be related to the data reference quality of each data attribute.
- In the third stage, combination and fusion of the best fitted data attributes, originates from different datasets, lead to information for the end-user based on the 'fitness for use' principle.

The model makes in all stage procedures use of the dynamic feedback mechanism.

Figure 1. Big Data Quality Assessment Model (BQDAM), Jellema, De Bakker, 2017



3 Example of Infofolio-application

By introducing, investigating, learning and working with the 'Big Data Quality Assessment Model', Infofolio follows the Theory of Change approach [Weiss, 1995]. This indicates that every case is used for further development of the model.

The following case is focussed on the data attribute 'volume of a building' and Infofolio use 8 different data sources with different coverage and quality. One data source 'pointclouds', derived from aerial photos, is discussed in more detail.

In stage 1 of the BDQAM the quality of the individual datasets is assessed. Aerial photos were good enough to deliver a Digital Elevation Model (DEM) with inverse distance interpolation (example of neighbourhood GI functionalities). Quality of the DEM was assessed with the measured points and indicated a 95% similarity. Volume of the building was calculated by vacuuming the building from the outside by comparing the points inside the footprint with the level of the surface.
In stage 2 we compare the calculated volume with the volume as registered in the other 7 data sources. Although the definition of the volume is different (often it excludes internal floors and walls, so it is net volume instead of gross volume) the regression analysis between the pointcloud volume and seven other data sources was 94%. Overall the reference quality was visualized as indicated in Figure 2 (from red: low to green: high quality).
In stage 3, the comparison and fusion of the 8 different data sources deliver a higher 'fitness for use' than the data of the data source 'pointclouds' alone.

Figure 2. Data reference quality of volume, Bakker, M. de et al, 2015

4 Conclusion

The proposed Big Data Quality Assessment Model delivers as shown by the case a better insight in the process of data quality assessment of Big Data. The approach with three different stages including the definition of data reference quality and 'fitness for use' makes the whole process more transparent for all stakeholders involved.

5 Acknowledgement

The authors like to thank Anne Muller for her review of our English.

6 References

Bakker, M. de, Voets D., Jellema, M., Bozelie, W. (2015) *Innovatie in kwaliteit door combinatie van puntenwolken en vastgoed-informatie*, Geo-Info 2015-02, pp 18 -20 The Netherlands.

By de, R.A. et al (2001) *Principles of Geographic Information Systems*, An introductory textbook, ITC, The Netherlands.

Cai, L., Zhu, Y. (2015) *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*, Data Science Journal 14.

DAMA UK Working Group *'Data Quality Dimensions'*, (2013) Defining Data Quality Dimensions, final paper.

Infofolio (2017) www.infofolio.nl

Jellema, M., Autar, A. (2015) *Hermes-model: Universal Model to Estimate The Rebuilding Costs of Houses*, FIG Working Week 2015, Sofia Bulgaria.

Jellema, M., Bakker, M. de (2017) *Quality Assessment of Big Data with GIS*, 20th AGILE-conference 2017, Wageningen The Netherlands.

Katal, A., Wazid, M., Goudar, R. (2013) *Big Data: Issues, Challenges, Tools and Good Practices*, Procedures of the 6th International Conference on Contemporary Computing, Noida India.

Klous, S., Wielaard, N. (2016) *Wij zijn Big Data*, Uitgeverij Business Contact, Amsterdam/Antwerpen.

Weiss, C. (1995) *Nothing as practical as good theory.*