Integrating Microdata on Higher Education Institutions (HEIs) with Bibliometric and Contextual Variables: A Data Quality Approach

Cinzia Daraio[1], Angelo Gentili[1] and Monica Scannapieco[2]

[1] *daraio@dis.uniroma1.it, angelo_gentili@hotmail.it*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *scannapi@istat.it*
Italian National Institute for Statistics (Istat), Rome (Italy)

An introduction on data quality

Data quality has been addressed in different research areas, mainly including statistics, management and computer science. The statistics researchers were the first to investigate some of the problems related to data quality by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 60s. The management research began at the beginning of the 80s; the focus was on how to control data manufacturing systems in order to detect and eliminate data quality problems. Only at the beginning of the 90s, computer science researchers began considering the data quality problem, specifically how to define measure and improve the quality of electronic data, stored in databases, data warehouses and legacy systems. Data quality has been defined as "fitness for use", with a specific emphasis on its subjective nature. Another definition for data quality is "the distance between the data views presented by an information system and the same data in the real world"; such a definition can be seen as an operational definition, although evaluating data quality on the basis of comparison with the real world is a very difficult task.

Data quality is well-recognized as a multidimensional concept including several distinct dimensions (Batini & Scannapieco, 2006) proposed in various contexts (Catarci & Scannapieco, 2002). A crucial dimension of data quality is data accuracy: it measures the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v is intended to represent. However, quality is more than simply data accuracy. Other significant dimensions play a role in the definition of the Data Quality concept, including completeness, consistency, and timeliness (i.e. degree of up-to-dateness), just to cite some significant ones.

Data Quality issues in data integration processes

In a data integration system, sources are typically characterized by various kinds of heterogeneities that can be generally classified into:

(i) Technological heterogeneities.
(ii) Schema-level heterogeneities.
(iii) Instance level heterogeneities.

Technological heterogeneities are due to the use of products by different providers, employed at various layers of an information and communication infrastructure.

Schema heterogeneities are principally caused by the use of (a) different data models, such as one source that adopts a relational data model and a different source that adopts a graph-based data model, and (b) different data representations, such as one source that stores addresses as one single field and another source that stores addresses with separate fields for street, civic number, and city. Schema level heterogeneities can be solved according to well-defined methods that harmonize data collected by the different sources with respect to a schema global to the whole data integration system. However, from a practical perspective, in order to make such harmonization possible it is also necessary to solve (iii) instance level heterogeneities, namely:

For overlapping data sources, same objects can be represented as different due to data quality errors. Hence, in order to resolve such conflicting representations, an object matching activity must be performed. Such activity should be as much automated as possible, especially in complex data integration systems (Zardetto, Scannapieco, Catarci, 2010).

For all sources, also those that are not overlapping, a quality control at instance-level is very useful in order to prevent the possible population of the data integration system with erroneous data. Depending on the specific types of data integration systems, such a quality control can be performed in different ways.

A Data Quality Approach to integrate HEIs microdata in a platform

For a platform supporting European Universities for Education, Research and Technology Studies, on the one hand, the lower level of disaggregation of data makes them more sensible and increases the chances of instance-level errors. On the other hand,

data collection is performed by integrating data already collected by statistical institutions by means of different statistical surveys or administrative data.

Hence, the quality control activity should have the following features:

1. It has to be applied on the overall collected data and cannot be applied to single processes producing data. Monitoring and control of processes producing data can be very useful to prevent quality problems, however, it cannot be applied to our case, due to the different nature of production processes and to the practical impossibility to revise such processes in a preventive fashion. This does not exclude of course the fact that feedbacks deriving from quality analysis could be used by organizations that produce data to revise their production processes.
2. A specific quality activity of outlier detection could be applied, by comparing data provided by "similar" sources on the same subject. Here, "similar" could mean, for instance, belonging to the same country and with analogous features such as the number of personnel. Data that are recognized as outlier by automated procedures should subsequently undergo a human analysis. This analysis can either explain the outlier on the basis of available context information, or it can recognize that the outlier is actually caused by quality problems. In this latter case, quality improvement actions must be engaged.

The following Table 1 illustrates the main sources of data which have been integrated to test the data quality approach proposed in the paper.

Figure 1 instead shows an example of outliers detection carried out through a systematic check against different distributions. The check has been done on the ratios given by number of publications divided by the number of academic staff, for all European universities in the sample.

Table 1. Main sources of data integrated

Source (link)	Description
ETER (www.eter.joanneum.at/ imdas-eter/) integrated with data from HESA for UK	Microdata on inputs outputs of higher education institutions in Europe.
Scimago Institutions Rankings (www.scimagoir.com)	Bibliometric data on scientific production and impact.
Eurostat (http://ec.europa.eu/eurostat)	Contextual factors, data at territorial level on economic and social development.

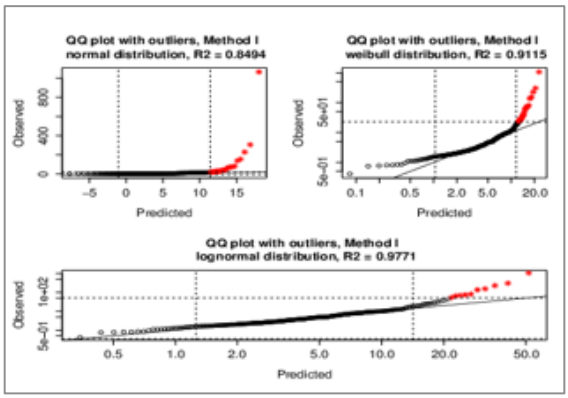

Figure 1. An example of outliers detection. Outliers are reported as stars in red: the graph top left shows outliers with respect to the normal distribution (worst fit, r-square=0.85), the one top right with respect to the Weibull distribution (r-square=0.91), the one below with respect to the lognormal distribution with the highest fit (r-square=0.98).

References

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies, and Techniques,* Springer, The Netherlands.

Catarci, T. & Scannapieco, M. (2002). Data Quality under the Computer Science Perspective. Journal of Archivi & Computer, 2.

Lepori, B., Daraio, C., Bonaccorsi, A., Daraio, A., Scannapieco, M., Gunnes, H., Hovdhaugen, E., Ploder, M. & D. Wagner-Schuster (2014), 'ETER Project, Handbook for Data Collection', Brussels, June.

Luwel, M. (2005). The use of input data in the performance analysis of R&D systems. In *Handbook of Quantitative Science and Technology Research* (pp. 315-338). Springer Netherlands.

Luwel, M. (2015), Heterogeneity of data in research assessment, in *Efficiency, Effectiveness and Impact of Research and Innovation*, Proceedings of the Workshop of the 20th February 2015, C. Daraio (ed), DIAG Sapienza University of Rome, Efesto Edizioni Rome.

Zardetto, D., Scannapieco, M., & Catarci, T. (2010). Effective Automated Object Matching. *Proceedings of the International Conference on Data Engineering (ICDE 2010)*.