Data Quality and Format for Remote Technologies

Alberto Sampaio, Gustavo Alves
CIETI/LABORIS
Instituto Superior de Engenharia do Porto (ISEP)
Porto, Portugal
{acs, gca}@isep.ipp.pt

Abstract: Quality of data is of paramount importance to the development of remote technologies. In this paper the authors discuss the need for research on data quality oriented to the field of remote technologies.

Information quality is a critical factor for activities in the information age [1]. In the case of remote technologies (RT), like remote labs, could be even more serious because, it is unquestionable that RT does not make sense without data that are of quality. It is know that poor data quality "reduces the value of the asset and can be expensive or impossible to correct" [2](p.10). Currently, data quality can be considered a reasonable well-studied subject, a claim supported by the existence of several textbooks (e.g., [3], [4]) and the proposal of standards directly related to data quality (e.g. ISO/IEC 25012 [5], ISO/IEC 11179 [6]).

However this state of practice and research is not true for the RT area. A very recent and comprehensive review of the state of the art in data quality assessment methodologies can be found in [7], but although recent and complete, there is no reference to data quality in the context of RT. RT infrastructures poses additional challenges, notably if we pretend to make any RT infrastructure available from anywhere to anyone, second, because there is a trend to consider RT infrastructures in group through some form of orchestration. Doing this implies to make available data from any RT system even to other RT infrastructures. This will be the case in a service oriented architecture, in the paradigm of RT as a service. All this means larges amounts of data and immense distinct sources of data produced in distinct ways and formats whose quality must be assured. All this can be worsened by the diversity of the several RT domains, like remote labs, telemedicine, and others. The amount of nodes involved would differ from other domains, like the traditional distributed information systems (e.g. see [7]) by number of sources and nature; from sensor networks, because these possess a limited number of distinct sensors; or even from geographical systems. More, the two latter can be, in some way, sub sources for EUREMOTE. At least this is a scalability issue, but if this is just a problem of scalability, is something that must be a subject of future research.

A significant part of RT data is expected to be unstructured or semi structured. It is known that research focusing on these kinds of data is limited and more research is needed [7][8]. Associated with data quality is the risk of poor quality obtained from such sources. For example, in the context of data mining "there is no formal methodological approach to dealing with the information quality risk for data mining in data warehouses" [1] (p.336). In the same paper is proposed a methodology to model quality risk but for a small

subset of data quality characteristics, that is, it is offered a partial solution and more research is also needed here.

User expectations and perceptions of data quality posed by the pervasive nature of the service-oriented architectures are expected and according [8] implying the need for more research. This is relevant for RT because of its service-oriented architecture and global computing availability.

To manipulate all the expect data, it will be easier using a format and tagging standard. Such format could act like metadata with implications in data quality. However, to the best of our knowledge in RT there is no common standard. However some initial steps are being taken (e.g., see, [10]). The format and tagging standard would make easier interoperability between RT services.

The aforementioned limited knowledge in all those areas of data quality offer prospects of research. Due the lack of data quality research for RT, notably about semi-structured and unstructured data, this task will have a string focus on the research of data qualities for RT. We plan to investigate data qualities for RT beginning from the data quality standard ISO/IEC 25012 and other models proposed by other researchers, both outside the RT area. After that, we wills start to develop a data quality system appropriate to the RT area. Nevertheless, we will try to capitalize from the knowledge about data quality generated in other areas, like, for example, geographical information systems, where some data and metadata standards have been produced (e.g., see [9]). This work will be done in parallel with the definition of a data tagging and format.

References

[1] Y. Su, J. Peng, and Z. Jin, "Modeling Information Quality Risk for Data Mining in Data Warehouses," *Human and Ecological Risk Assessment: An International Journal*, vol. 15, 2009, p. 332.

[2] G.C. Simsion and G.C. Witt, *Data modeling essentials*, Morgan Kaufmann, 2005.

[3] R.Y. Wang, M. Ziad, and Y.W. Lee, *Data Quality*, 2001.

[4] C. Batini and M. Scannapieco, *Data quality*, Springer, 2006.

[5] "ISO/IEC FCD 25012:2008 - Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model."

[6] "ISO/IEC 11179-1:2004 - Information technology -- Metadata registries (MDR) -- Part 1: Framework."

[7] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, 2009, pp. 1-52.

[8] S.E. Madnick, R.Y. Wang, Y.W. Lee, and H. Zhu, "Overview and Framework for Data and Information Quality Research," *J. Data and Information Quality*, vol. 1, 2009, pp. 1-22.

[9] R. Devillers and R. Jeansoulin, *Fundamentals of Spatial Data Quality*, Wiley-ISTE, 2006.

[10] Maier, C., Niederstätter, M., Lab2go – A Repository to Locate Online Laboratories, International Journal of Online Engineering (iJOE), Vol 6, No 1, 2010.