

Quality Issues of CRIS Data

Otmane Azeroual^a, Joachim Schöpfel^b

^a German Center for Higher Education Research and Science Studies (DZHW), Schützenstraße 6a, 10117 Berlin, Germany

Azeroual@dzhw.eu

^b GERiCO laboratory, University of Lille, 59650 Villeneuve-d'Ascq, France

Joachim.schopfel@univ-lille3.fr

Track III. Advances in research information technology

Research paper (short submission)

Extended abstract

In recent years, current research information systems have become an integral part of university IT landscapes, and in this regard, their importance in research has greatly increased. Through the use of CRIS, scientific institutions can provide a current overview of their research activities, collect, process and manage information about their scientific activities, projects and output as well as integrate them into their web presence. Furthermore, CRIS can contribute to rationalize and optimize academic research activities, through highly efficient procedures and added value information.

In the context of the scientific Big Data, research and development studies often focus on the “3 Vs”, i.e. volume, velocity and variety. Our paper will address the 4th and 5th Vs, ie data variability (inconsistency), and data veracity (quality) as a specific problem for research information systems. Here, we distinguish between the quality of the system and the quality of the content, in particular of the data input.

Data trustworthiness plays an important role in transforming the data into meaningful information. Therefore, the success or failure of a CRIS in a scientific institution is significantly related to the quality of the data input available as a basis for the CRIS applications. The most beautiful business intelligence (BI) tools, such as reporting, are worthless when incorrect data is entered.

The collection, transmission and integration of research information into CRIS can result in different data errors (such as missing values, duplicates, spelling errors, incorrect formatting, inconsistency, etc.), with a variety of negative effects on information quality. Small mistakes can make data unusable within universities and research institutions, and corrupted data can have fatal consequences. The sooner quality issues are identified and remedied, the better. A one-time cleanup is not sufficient, data must be continuously curated and maintained.

Quality has always been an issue of the CRIS community, for the industry (service provider) as well as for the academic institutions (user). A selective overview reveals about 25 representative papers presented at euroCRIS meetings between 2002 and 2017. Together with some articles on CRIS and data quality published in different academic journals, they provide insight in three main dimensions of this topic:

Concepts of data quality: what is information quality? what is reliable data? what are criteria for the assessment of quality? how to assess quality? Sometimes these questions are related to a risk analysis. Based on these studies and on older works on data and information quality, we'll try to propose a new concept of data quality in the environment of research information systems.

Quality approach: which kind of control and monitoring of data quality? which kind of quality assurance procedure? which data quality strategy? For instance, one interesting approach to data quality has been presented at CRIS2016 by Ivan Nevolin from the Russian Academy of Sciences, ie crowdsourcing. We'll try to develop some guiding principles based on these studies.

Standards: standards are crucial for data quality, and many papers focus on metadata, formats, identifiers, data definitions and so on. We'll provide a short overview of the main aspects, including recent work from euroCRIS, EUNIS and CASRAI. Here, we will also make a link to the European FAIR research data policy and its impact on CRIS.

In order to detect and correct errors of data sources and input at an early stage, the scientific institutions must determine the new techniques or methods, such as data cleansing, in order to provide an adequate and reliable decision-making basis in a CRIS and to increase confidence in the available data. With the help of these new techniques, the scientific institutions can evaluate their research information and provide information about their quality. Moreover, they can assess the dependencies, redundancies and anomalies between data fields and improve the monitoring and correction within their CRIS.

Our paper identifies data quality issues in research information systems and introduces new techniques and methods of data cleansing that enable organizations to improve the quality of their research information. We will provide an overview of relevant studies, especially from former CRIS conferences and meetings, and we will apply concepts (and assess their relevance) from the information sciences to the CRIS data quality issues, eg metrics such as accuracy, objectivity, relevance, timeliness, interpretability, compatibility or accessibility.

We will also conduct interviews with a panel of experts and stakeholders from the CRIS industry (publishers, IT companies), from institutional CRIS customers and universities, research institutes (technical staff) from different EU member states to assess their quality issues, their risk analysis and their approach to prevent, detect, monitor and correct data quality problems. In particular, we will address five topics:

- Quality awareness in general
- Quality of data input
- Control (monitoring) of data input (e.g., names, affiliations, metadata of research output, missing data...).
- Correction (cleansing) of data input, for instance in case of data inconsistency (by humans and/or machines/algorithms)
- Impact of CRIS data monitoring on institutional data sources (HR database...)

Our paper presents preliminary results of a research project in progress. The paper's originality is its empirical character based on a representative state of the art and a conceptual analysis of the issue, and its multidisciplinary perspective combining computer and information sciences. Based on our conceptual analysis and empirical results, the paper will conclude with some recommendations for further CRIS research and data management, including good practice and new technologies.

Keywords

Current Research Information System, CRIS, Research Information System, RIS, Research Information, Data Sources, Quality Issues, Data Quality, Data Cleansing, Data Monitoring

References

- Anderson, J. C., Rungtusanatham, M. and Schroeder, R. G. (1994). A theory of quality management underlying the deming management method. *The Academy of Management Review*, 19(3):472–509, 1994.
- Azeroual, O. and Abuosba, M. (2017). Improving the Data Quality in the Research Information Systems. *International Journal of Computer Science and Information Security*, 15(11): 82-86, November 2017.

- Azeroual, O., Saake G. and Abuosba, M. (2018). Data Quality Measures and Data Cleansing for Research Information Systems. *Journal of Digital Information Management*, 16(1): 12-21, February 2018.
- Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):1–52, July 2009.
- Berkhoff, K., Ebeling, B., and Lübke, S. (2012). Integrating research information into a software for higher education administration – benefits for data quality and accessibility. In *CRIS2012: 11th International Conference on Current Research Information Systems*, Prague, June 6-9, 2012.
- Biesenbender, S. and Hornbostel, S. (2016). The research core dataset for the German science system: challenges, processes and principles of a contested standardization project. *Scientometrics*, 106(2):837-847.
- Boldrini, E., Luzi, D., Nativi, S., and Pecoraro, F. (2014). Integrating CERIF entities in a multidisciplinary e-infrastructure for environmental research data. In *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014.
- Chudlarský, T. and Dvořák, J. (2012). A national CRIS infrastructure as the cornerstone of transparency in the research domain. In *CRIS2012: 11th International Conference on Current Research Information Systems*, Prague, June 6-9, 2012.
- English, L. P. (2009). Information quality applied: *best practices for improving business information, processes, and systems*. Wiley, Hoboken NJ.
- Evans, B., Druken, K., Wang, J., Yang, R., Richards, C., and Wyborn, L. (2017). A data quality strategy to enable FAIR, programmatic access across large, diverse data collections for high performance data analysis. *Informatics*, 4(4):45+.
- Fondermann, P. and Van der Togt, P. (2016). How Wageningen University and research center managed to influence researchers publishing behaviour towards more quality, impact and visibility. In *CRIS2016: 13th International Conference on Current Research Information Systems*, St Andrews, June 9-11, 2016.
- Hahnen, H. and Güdler, J. (2008). Quality is the product is quality – information management as a closed-loop process. In *CRIS2008: 9th International Conference on Current Research Information Systems*, Maribor, June 5-7, 2008.
- Hubermann, M. A. and Miles, M. B. (1998). Data management and analysis methods. In N. K. Denzin and Y. S. Lincoln, editors, *Collecting and Interpreting Qualitative Materials*. Sage Publications Inc., Thousand Oaks, CA, 3rd edition, February 1998.
- Jörg, B., Höllrigl, T., and Sicilia, M. A. (2012). Entities and identities in research information systems. In *CRIS2012: 11th International Conference on Current Research Information Systems*, Prague, June 6-9, 2012.
- Ilva, J. (2016). Towards reliable data - counting the finnish open access publications. In *CRIS2016: 13th International Conference on Current Research Information Systems*, St Andrews, June 9-11, 2016.
- Krcmar, H. (2009). Informationsmanagement. Springer, 5 edition, November 2009.
- Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & Management*, 40(2):133–146, December 2002.
- Leser, U. and Naumann, F. (2007). Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen. Dpunkt Verlag, 1. a. edition, 2007.

- Lingjærde, G. C. and Sjøgren, A. (2008). Quality assurance in the research documentation system frida. In *CRIS2008: 9th International Conference on Current Research Information Systems*, Maribor, June 5-7, 2008.
- Madnick, S. E., Wang, R. Y., Lee, Y. W. and Zhu, H. (2009). Overview and framework for data and information quality research. *J. Data and Information Quality*, 1(1):1–22, June 2009.
- Magnani, M. and Montesi, D. (2010). A survey on uncertainty management in data integration. *J. Data and Information Quality*, 2(1):1–33, July 2010.
- Miles, M. B. and Huberman, A. M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*(2nd Edition). SAGE Publications, Inc, 2nd edition, January 1994.
- Miller, H. (1996). The multiple dimensions of information quality. *Information Systems Management*, 13(2):79-82.
- Naumann, F. (2007). Datenqualität. *Informatik-Spektrum*, 30(1):27–31, February 2007.
- Naumann, F. and Rolker, C. (2000). Assessment methods for information quality criteria. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 148–162, 2000.
- Nevolin, I. (2016). Crowdsourcing opportunities for research information systems. In *CRIS2016: 13th International Conference on Current Research Information Systems*, St Andrews, June 9-11, 2016.
- Pipino, L. L., Lee, Y. W. and Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4):211–218, 2002.
- Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.
- Raman, V. and Hellerstein, J. M. (2001). Potter's wheel: An interactive data cleaning system. In *The VLDB Journal*, pages 381–390, 2001.
- Ribeiro, L., de Castro, P., and Mennielli, M. (2016). *EUNIS – EUROCRIS joint survey on CRIS and IR*. Final report. ERAI EUNIS Research and Analysis Initiative, Paris.
- Stempfhuber, M. (2008). Information quality in the context of CRIS and CERIF. In *CRIS2008: 9th International Conference on Current Research Information Systems*, Maribor, June 5-7, 2008.
- Tolstoy, T. (2016). Swepub analysis - serving high quality repository data and more using linked data technologies. In *CRIS2016: 13th International Conference on Current Research Information Systems*, St Andrews, June 9-11, 2016.
- Van den Berghe, S. and Van Gaeveren, K. (2016). Data quality assessment and improvement: a vrije universiteit brussel case study. In *CRIS2016: 13th International Conference on Current Research Information Systems*, St Andrews, June 9-11, 2016.
- Wang, R. Y., Kon, H. B., and Madnick, S. E. (1993). Data quality requirements analysis and modeling. In *Data Engineering, 1993. Proceedings. Ninth International Conference on*, pages 670–677. IEEE, April 1993.
- Wang, R. Y., Storey, V. C. and Firth, C. P. (1995). A framework for analysis of data quality research. *Knowledge and Data Engineering, IEEE Transactions on*, 7(4):623–640, 1995.
- Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5-33.

Authors Profile

Otmane Azeroual, is researcher at the German Institute for Higher Education Research and Science Studies (DZHW) in Berlin. After studying Business Information Systems at the University of Applied Sciences Berlin (HTW), he began his PhD in Computer Science at the Institute for Technical and Business Information Systems (ITI), Database Research Group of the Otto-von- Guericke-University Magdeburg and at the Department of Computer Science and Engineering of the University of Applied Sciences (HTW) Berlin. His research interest is in the area of Information Systems, Database Systems, Data Quality Management, Business Intelligence, Big Data, IT- Security, Cloud Data Management, Open Data, Project Management and Industry 4.0.

Joachim Schöpfel, is senior lecturer in Library and Information Sciences at the University of Lille (France), researcher at the GERiCO laboratory and consultant at the Ourouk consulting firm. He was Manager of the INIST (CNRS) scientific library from 1999 to 2008, head of the University of Lille department of information sciences from 2009 to 2012 and director of the French Atelier National de Reproduction des Thèses (ANRT) from 2012 to 2017. He teaches library marketing, auditing, valorization and digitization of cultural heritage collections, intellectual property and information science. His research interests are scientific information and communication, especially open access and open science, research data and grey literature. He is member of euroCRIS, of the NDLTD board of directors and of Knowledge Exchange.