Introduction to Record-Linkage

Rainer Schnell Severin V. Weiand

13.4.2026, BERD-Workshop

Abstract

Record linkage refers to the process of identifying and linking records from multiple data sources that correspond to the same underlying entity – for example, matching individuals across distinct administrative databases. By enabling the integration of separate datasets of the same persons or organisations, record linkage substantially expands the analytical potential of available data. Applications include the construction of historical panels, the validation of survey responses by using process-generated data, and the estimation of the size of rare populations.

This workshop provides a comprehensive introduction to record linkage for practitioners across diverse scientific disciplines, including medicine, the social sciences, and economics. It addresses the fundamental methodological challenges of record linkage, particularly data quality, classification error, and privacy protection. Participants will gain a practical understanding of these challenges and the conceptual foundations required to address them effectively.

Special emphasis will be placed on privacy-preserving record linkage (PPRL), which aims to enable accurate linkage while safeguarding sensitive personal information. The workshop concludes with a hands-on computational demonstration in R that illustrates selected approaches. These include probabilistic record linkage based on the Fellegi-Sunter model and privacy-preserving techniques employing Bloom filters.

Learning Objectives

- Record linkage fundamentals: Participants will understand the essential concepts of record linkage, including its purpose for using data from various sources to identify the same entity.
- Explore diverse applications: Attendees will discover how record linkage is applied across multiple scientific disciplines to gain a wider understanding of the analytical use of merged data.
- Identify data quality challenges: Participants will learn to recognise common data quality issues in record linkage and understand how these challenges can influence statistical outcomes, including differential linkage bias.

Learning Objectives

- Understand privacy considerations: The workshop will equip attendees with knowledge of the privacy problems resulting from record linkage. The focus will be on the technical solutions to avoid these problems.
- Record linkage methods: Participants will develop elementary skills in deterministic and probabilistic record linkage, as well as PPRL using Bloom filters.
- Conducting record linkage: The ability to conduct simple linkages using these methods in R.

References

Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer. https://doi.org/10.1007/978-3-642-31164-2

Christen, P., Ranbaduge, T., & Schnell, R. (2020). Linking sensitive data: Methods and techniques for practical privacy-preserving information sharing. Springer. https://doi.org/10.1007/978-3-030-59706-1

Herzog, T. N., Scheuren, F., & Winkler, W. E. (2007). Data Quality and Record Linkage Techniques. Springer. https://doi.org/10.1007/0-387-69505-2

Schedule

09:00–09:45 Block 1: Introduction to Record Linkage

09:45–10:00 Break

10:00–10:45 Block 2: The Record Linkage Process

10:45–11:00 Break

11:00–11:45 Block 3: Data Quality and Privacy-Preserving Record Linkage

11:45–12:00 Break

12:00–13:00 Block 4: Practical Demonstration

Lecturers

Prof.em. Dr. Rainer Schnell is a senior professor at the University of Duisburg-Essen. Previously, he was a professor at the University of Konstanz and the City University London. He has more than 20 years of experience in record linkage and is one of the leading experts in Privacy-Preserving Record Linkage (PPRL).

Severin V. Weiand is a doctoral candidate at the University of Duisburg-Essen, working in his doctoral thesis on large-scale record linkage. He holds a BA in computer science, a BA in sociology and an MA in behavioural data science. Over the past five years, he has worked with several federal agencies on data linking issues.