ADR UK is hosting a two-day on-campus workshop

Workshop on privacy-preserving record linkage (PPRL)

3 and 4 June 2025

The event will be run by ADR Scotland and held in Edinburgh.

Overview:

Record linkage is the process of identifying records that refer to the same real-world entities (often individuals) across two or more databases. The main aspect that makes record linkage challenging is the lack of unique entity identifiers (such as NHS numbers) available in all the databases to be linked. Therefore, the personal details of individuals, such as their names, addresses, and dates of birth (commonly known as quasi-identifiers) are used to compare and link records.

Such an approach will, however, be challenged by data quality (erroneous, missing, and out-of-date values) and privacy regulations (which limit or prohibit the exchange of personal data). To overcome the second challenge, **privacy-preserving record linkage** (PPRL) has seen increasing interest across a variety of domains in the past three decades.

In this two-day workshop, participants will learn about the concepts, methods, and challenges of PPRL, and hear from experts in the domain about state-of-the-art research, practical real-world applications of PPRL, as well as future directions.

The workshop will run over two days, where on the first day presentations will focus on the general concepts of the PPRL process, describe key PPRL protocols and techniques, and highlight the limitations and vulnerabilities of existing PPRL methods. The second day will then be a mix of presentations by researchers and practitioners from organisations who are employing PPRL methods. The workshop will also include practical sessions where participants will be able to work with simple Python-based PPRL programs (provided to participants) to explore PPRL methods using small publicly available example databases.

This PPRL workshop aims to provide:

- An introduction to record linkage and PPRL
- Highlight the potential challenges with record linkage and PPRL
- Demonstrate how privacy-preserving record linkage can resolve such challenges
- Hear from experts who will discuss practical real-world applications of PPRL and future directions of PPRL research
- Computer-based lab exercises - using PPRL methods on sample data sets

Timetable of the Workshop

Day One – 3 June:

The first day will have **four tutorial style presentations**, focusing on the general concepts of the PPRL process, describing key PPRL protocols and techniques, and highlighting the limitations and vulnerabilities of existing PPRL methods.

These presentations will be given by Prof Peter Christen and Prof Rainer Schnell.

09:30 onwards:	**Registration and coffee**, all participants expected to arrive by 10 am
10:15 to 10:30:	Welcome, outline of the workshop and program of the first day
10:30 to 11:30:	**Presentation 1**: Introduction to record linkage, the record linkage process, PPRL overview, scenarios, and challenges
11:30 to 11:45:	Short comfort break
11:45 to 12:45:	**Presentation 2**: PPRL methods and techniques, including key building blocks of PPRL, and selected popular PPRL methods such as Bloom filter and match-key based PPRL
12:45 to 13:15:	**Discussion and question session** for topics discussed so far
13:15 to 14:00:	**Lunch**
14:00 to 15:00:	**Presentation 3**: Vulnerabilities of PPRL methods, attacks on PPRL, and hardening techniques to overcome such attacks
15:00 to 15:30:	Coffee break
15:30 to 16:30:	**Presentation 4**: Practical applications of PPRL, and relevant real-world aspects of PPRL, considerations outside of core technical PPRL methods relevant for real-world systems.
16:30 to 17:00:	**Discussion and question session** for topics covered on the first day

We plan to limit the day 1 attendance to around 50 participants.

Day Two – 4 June:

On the second day of the workshop, in the morning we plan for two parallel streams. The first will be practical sessions, where participants will work with simple Python-based PPRL programs to explore PPRL methods using small publicly available example data sets. Or they can select Stream 2, where there will be a mix of presentations by researchers and practitioners from organisations who are employing PPRL methods.

We end the workshop with a combined final round table and Q & A discussions session, running from 14:00 to 15:00, followed by short closing remarks by Prof Chris Dibben.

Stream 1: Practical PPRL:

Practical PPRL sessions, facilitated by Dr Charini Nanayakkara (ANU and UoE) and Ms Sumayya Ziayad (ANU).

Based on computer-based lab exercises where participants are given Python skeleton programs and work through a set of conceptual and programming tasks to become familiar with selected PPRL methods (hash encoding and Bloom filter encoding) on sample data sets.

9:00 to 9:30:	**Registration and Welcome to Stream 1 participants**
9:30 to 11:00:	**Session 1**
11:00 to 11:30:	Coffee break
11:30 to 13:00:	**Session 2**
13:00 to 14:00:	Lunch
14:00 to 15:00:	**Combined final round table and Q & A discussion session**, followed by short closing remarks by Prof Chris Dibben

We plan to limit the attendance of this stream to 20 participants.

Stream 2: Real-world aspects of PPRL:

We plan a mix of talks by researchers and practitioners from organisations who are employing PPRL method, followed by discussion and Q & A session. For abstracts and speaker bios see below.

9:00 to 9:30:	**Registration and Welcome to Stream 2 participant**
9:30 to 11:00:	**Session 1** (1) **Keynote**: Australian examples of privacy-preserving record linkage using Bloom filters (Prof James Boyd, La Trobe University, Melbourne) (2) PPRL at the ONS: Enabling Quality Assurance (Josie Plachta and Leah Maizey, ONS Methodology Data Linkage team)
11:00 to 11:30:	Coffee break
11:30 to 13:00:	**Session 2** (1) Linkage within SeRP and SAIL: From trusted third parties to linkage as a service (Dr Mike Edwards, SAIL / Swansea University) (2) Practical Requirements for PPRL (Prof Rainer Schnell) (3) Beyond Bloom Filters: A single parameter method for secure privacy-preserving record linkage (Sumayya Ziyad, Peter Christen, Anushka Vidanage, Charini Nanayakkara, and Rainer Schnell)
13:00 to 14:00:	Lunch
14:00 to 15:00:	**Combined final round table and Q & A discussion session**, followed by short closing remarks by Prof Chris Dibben

Speakers:

Prof Peter Christen (Univ of Edinburgh and Australian National University); Peter Christen is the Research Lead on the Scottish Historic Population Platform (SHiPP) project, run at the Scottish Centre for Administrative Data Research (SCADR) at the University of Edinburgh. He is also a Professor at the School of Computing at the Australian National University in Canberra. Peter is a world-leading expert in record linkage with over 20 years' experience in working with administrative data. He has over 200 publications in the area of data science, including the two books "Data Matching" in 2012 and "Linking Sensitive Data" (co-authored with Thilina Ranbaduge and Rainer Schnell) in 2020. Peter is an award winning university lecturer who has been teaching in the area of data science since 2002. He has developed multiple large courses on topics such as Data Mining and Data Wrangling, and given workshops and tutorials on privacy aspects of record linkage since 2008.

Prof Rainer Schnell: Rainer Schnell is Professor for Research Methodology at the University of Duisburg-Essen. He has been working on methods for data collection, editing and linking population-covering datasets for 30 years. His main area of expertise is using computational statistical methods to identify and reduce human behaviour's impact on large-scale data collection efforts. During the last twenty years, Rainer has worked for National Statistical Institutes in four European countries and was involved in designing and executing national and cross-border record-linkage operations. He has written two textbooks on research methodology, a monograph on nonresponse in surveys, a textbook on "Linking Sensitive Data" (co-authored with Peter Christen and Thilina Ranbaduge) and about 100 papers and technical reports on record linkage.

Prof James Boyd: (La Trobe University, Australia)**:** Professor James Boyd is the inaugural Chair of Digital Health at La Trobe University. He has a research background and is an international data linkage expert leading La Trobe's Digital Health strategy. The digital health program at La Trobe University aims to address limitations and inefficiencies in the healthcare system resulting from the lack of 'joined-up' information, evidence and knowledge. Professor Boyd has over 25 years of experience working with large, linked population-based health administrative datasets to produce national epidemiological and management information, assisting in monitoring and evaluating health service performance.

Dr Charini Nanayakkara: Charini is a Postdoctoral Research Fellow at the Australian National University (ANU) in the School of Computing, where she conducts research on record linkage techniques for integrating large-scale population data. She contributes to the design and implementation aspects of novel advanced record linkage methods in the Scottish Historic Population Platform (SHiPP) project, which aims to integrate millions of Scottish civil registration certificates (births, deaths and marriages) between 1855 and 1973. She is also a co-convenor and lecturer of the Data Wrangling course offered at the ANU, teaching cohorts of over 200 undergraduate and Master students. Charini will be conducting the practical PPRL session on day two.

Ms Sumayya Ziyad: Sumayya is a PhD student at the School of Computing at the Australian National University. Her research focuses on developing efficient privacy-preserving record linkage techniques with provable privacy guarantees. She is also deeply interested in the fairness and bias of such techniques, working towards developing fair linkage methods. Sumayya has also conducted tutorials for the Data Wrangling course at ANU, both for undergraduate and Master students, guiding students in implementing record linkage solutions. Sumayya will be supporting the practical PPRL session on day two.

Josie Plachta and Leah Maizey (Office of National Statistics) Josie Plachta and Leah Maizey co-lead the Methodology Data Linkage team. While Josie is responsible for leading projects that require bespoke application of data linkage methods, or the development of new methods for linkage, Leah is responsible for the application and development of methods involved in quality assuring linked data. This team works regularly as a trusted-third-party in linking sensitive data for projects of significant national importance such as the censuses and providing datasets to inform COVID-19 decision making. They have also explored privacy preserving record linkages, including devising a linkage method that enables linking of hashed data and investigation. Throughout all of this work they champion data linkage quality by incorporating it into their work and providing advice to enable others.

Dr Mike Edwards (SAIL / Swansea University) is Principal Data Linkage Architect at Secure e-Research Platform UK (SeRP), based at Swansea University. He is responsible for the design, development, and deployment of linkage solutions within the SeRP Linkage Technology arm. His main research interests are in graph-based learning, record linkage, and the effective communication of linkage insights with data owners and users.

Abstracts for Day two Stream 2 presentations:

Prof James Boyd: Australian examples of privacy-preserving record linkage using Bloom filters

Abstract: This presentation examines the application of Privacy-Preserving Record Linkage (PPRL) using Bloom filters in sensitive Australian data linkage environments. With increasing organisational awareness of PII risks, this presentation outlines case studies demonstrating how PPRL (using Bloom Filters) facilitated secure data linkage between diverse entities, including state and Commonwealth agencies, healthcare providers, and criminal justice systems. The presentation will analyse the defining elements, risks, solutions, quality, and performance of these applications, highlighting challenges and opportunities for future improvement. The findings highlight the role of PPRL in enabling secure data linkage in risk-sensitive contexts, while emphasising the need for ongoing adaptation to evolving privacy demands.

Josie Plachta and **Leah Maizey**: PPRL at the ONS: Enabling Quality Assurance

Abstract: Linkage is a keystone of the UK Office For National Statistics in its production of data to support better government decisions. As a result, linkage must be performed to a high quality standard, and that quality must be understood to ensure that analysts can be confident in their results. Given the importance of Clerical Matching to understanding linkage quality, Privacy Preserving Record Linkage makes measuring linked data quality a challenge. At ONS, two main methods of PPRL have been employed: The Trusted Third Party System and a collaboration between ONS and DWP to share Hashed Data to be linked by the Bespoke Derive and Conquer algorithm alongside a sample of in the clear data to enable clerical for estimating quality. Both of these methods include quality assessment. This talk will cover the importance of data linkage quality, and how these two approaches allow us to link highly sensitive data securely while still understanding the quality of the linked data, including reference to some ONS case studies.

Prof Rainer Schnell: Practical Requirements for PPRL

Abstract: In 2011, we established the first academic data linkage facility (the German Record Linkage Centre) within the Research Data Centre of the Social Security Administration. During the last decade, we have conducted numerous linkages for third parties. The linked data originated from various sources such as individual, household or establishment surveys, administrative data, commercial company data and publicly available data. Furthermore, some administrations asked for the simulation of future linkages. Finally, we advised some national linkage projects on required identifiers and parameter settings. Based on this experience, this presentation will discuss the practical requirements for conducting a PPRL project. These requirements are 1) the existence of a research problem relevant to PPRL, 2) the availability of suitable databases, 3) the legal access to these databases, 4) the amount and data quality of identifiers and 5) the scalability of the PPRL technique. Real-world examples will illustrate each requirement. Knowing these requirements and communicating these to clients will help avoid unnecessary efforts and concentrate on technical issues, which can usually be solved.

Dr Mike Edwards: Linkage within SeRP and SAIL: From trusted third parties to linkage as a service

Abstract: Record linkage within SAIL is undertaken through collaboration with a Trusted Third Party (TTP) via a split-file process, with personally identifiable information kept separate from the rest of the record information to meet information governance requirements and minimise re-identification risks. Linkage is handled by the TTP, producing linking tables which are used in provisioning linked cohorts for research across a wide range of administrative and healthcare domains. A new service from Secure eResearch Platform (SeRP) offers linkage as an in-house service without the need for a third party, and with it comes the need to consider approaches to maintain a separation between identifiable data and the data researchers utilise in their research. This talk will provide an overview of the current approach to linkage within SAIL, discuss movements to providing a new linkage service, explore the intricacies between a TTP setup and an in-house service, and stimulate conversation around how privacy preserving record linkage can be utilised to provide an appropriate linkage solution.

Prof Peter Christen: Beyond Bloom Filters: A single parameter method for secure privacy-preserving record linkage

Abstract: Record linkage is the process of matching records that refer to the same entities (often people) across databases. In applications such as health research or government services, the databases to be linked are often sensitive and cannot be shared between organisations. Privacy-preserving record linkage (PPRL) aims to overcome this challenge by facilitating the comparison of encoded or encrypted records without having to share sensitive data. Most existing PPRL techniques are based on heuristics and they have limitations in the privacy protection they offer, such as being vulnerable to certain cryptanalysis attacks. Furthermore, existing PPRL methods have multiple parameters, which, if not set properly by the user, can result in sub-optimal linkage quality and reduced privacy protection. We present a novel PPRL method that uses random reference q-gram sets to generate bit-arrays that represent sensitive values. Our method has a single parameter to be set by the user that trades scalability with linkage quality and privacy protection. All other parameters are either data-driven or have strong bounds based on this user parameter. We conceptually analyse our method and conduct experiments on multiple databases. The results demonstrate that our method provides high linkage quality and strong privacy protection while being scalable to link very large databases.