International Journal of Population Data Science

Journal Website: www.ijpds.org

An Introduction to DLforum – An online discussion forum for data linkage researchers and practitioners

https://dmm.anu.edu.au/DLforum/

Christen, P[1*], Ranbaduge, T[1], and Vatsalan, D[1, 2]

Submission History

Submitted:	22/09/2017
Accepted:	14/12/2017
Published:	20/02/2018

[1]Research School of Computer Science, The Australian National University, Canberra, Australia
[2]Data61, CSIRO, Eveleigh, NSW, Australia

Abstract

Data linkage, the process of identifying records that refer to the same entities across databases, is a crucial component of Population Data Science. Data linkage has a history going back over fifty years with many different methods and techniques being developed in various disciplines including computer science, statistics, and health informatics. Data linkage researchers and practitioners are commonly only familiar with methods and techniques that have been developed or are used in their own discipline, and they often only follow research that is being published at venues in their own discipline. There is currently no single online resource that allows data linkage researchers and practitioners across different disciplines to exchange ideas, post questions, or advertise new publications, software, open positions, or upcoming conferences and workshops. This leads to a communication gap in the multi-disciplinary field of data linkage. We aim to address this gap with the *DLforum*, a public online discussion forum for data linkage. *DLforum* contains several discussion areas, including publication announcements, resources (software and datasets), information about upcoming conferences and workshops, job opportunities, and general questions related to data linkage. The forum includes a moderation process where all registered users can post content and reply to posts by other users. We anticipate that the number of users registered and the amount of content posted in the forum will show that such an online forum is of value to data linkage researchers and practitioners from different disciplines to effectively communicate and exchange their knowledge, and thus form an online community of practice. In this paper we describe the methods of developing the *DLforum*, its structure and content, and our plan on how to evaluate the forum. The *DLforum* is freely available at: https://dmm.anu.edu.au/DLforum/

Introduction

Data linkage is the increasingly important topic of how to identify and link records about entities across several databases [1,2,3]. The entities to be linked in most application domains refer to people (such as patients, customers, tax payers, travellers, and so on), while in other applications entities such as publications from bibliographic databases [4] or consumer products (from online shopping sites) need to be linked [5].

Data linkage has a long history going back to Halbert Dunn [6] who used the term record linkage to describe the idea of assembling a book of life for individuals. Each such book would start with a birth record and end with a death record, and in-between would contain marriage and divorce records as well as records about an individual's contacts with the health and social security systems. Dunn realised that having such books for all individuals in a population would provide a wealth of information that would allow improved national statistics, better planning of services, and facilitate research studies in various domains. Over the past six decades much research and development has been conducted in the area of data linkage, notably the work by Newcombe and colleagues in the 1950s and 1960s [7,8] on computer-based linkage which was followed by the seminal work by Ivan Fellegi and Alan Sunter on *probabilistic record linkage* [9]. Their approach is still the basis of many current data linkage systems used in practice.

In more recent times, with the advent of increasingly large databases being collected by businesses and governments [10], and research in various disciplines becoming increasingly data-oriented, novel methods and techniques for data linkage have been developed in a variety of disciplines, especially in the statistical and computing sciences (see, for example, Herzog et al. [3], Christen [1], and Dong and Srivastava [10]). While applications of data linkage in different disciplines have to deal with specific challenges, overall most data linkage applications encounter the following three main challenges [1]:

1. **Data quality**, because often no unique entity identifiers (such as person identifiers) are available in all the databases to be linked. Therefore quasi-identifying fields

*Corresponding Author:
Email Address: peter.christen@anu.edu.au (P Christen)

http://dx.doi.org/10.23889/ijpds.v3i1.420

or attributes (such as names, addresses, and dates of birth) need to be used for the linkage. However, the values in these fields can contain errors and variations, be out-of-date, or be missing.

2. **Scalability** to linking larger databases as well as linking data from an increasing number of data sources, because the amount of data being collected is steadily increasing. Additionally, advanced population studies often require information from diverse organisations to be linked to investigate today's complex challenges [11].

3. **Privacy and confidentiality**, because the data to be linked is commonly about people alive today, and the growing concern by the public about the use of their personal information limits how sensitive personal data can be linked across organisations (see, for example, Hall and Fienberg [12] and Vatsalan et al. [13]).

While the challenges of conducting data linkage are similar across disciplines (including computer science, statistics, and the health and social sciences) and application areas (in both the private and public sectors), different solutions have been and are being developed to tackle these challenges. Researchers in one discipline are often not aware of solutions to their problems that have been developed in other disciplines. This can lead to the re-development (and re-publication) of similar ideas in different disciplines. The wide range of research and application areas of data linkage is illustrated by the breadth of three recently published books on this topic: Herzog et al. [3] address data linkage from a statistical perspective, Christen [1] mostly covers advances in data linkage from computer science, and the edited book by Harron et al. [2] includes authors from national statistics, the social and health sciences, as well as computer science. The lack of exchange of ideas, methods and techniques, and best practice approaches across disciplines, hampers the use of data linkage in many applications and potentially leads to non-optimal outcomes of data linkage projects.

The International Population Data Linkage Network (IPDLN, http://www.ipdln.org) and its bi-annual conferences, and the new International Journal of Population Data Science (IJPDS, https://ijpds.org), are fantastic opportunities to bring researchers and practitioners working in data linkage together. However, what is currently missing is an online resource that allows more interactive ongoing discussions as well as announcements of any new advances in the field of data linkage.

The idea of an online forum for data linkage grew out of discussions at the workshop "Data Linkage: Techniques, Challenges, and Applications" (https://www.newton.ac.uk/event/dlaw02), held in September 2016 at the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge as part of the "Data Linkage and Anonymisation" programme (https://www.newton.ac.uk/event/dla). Researchers and practitioners who attended the workshop, with a diverse background including public health, government services, the social sciences, as well as from businesses, highlighted that interaction across the disciplines is highly important to facilitate cross-disciplinary learning and to tackle the big challenges of data linkage. The workshop discussions identified that having mechanisms in place that allow ongoing interactions between researchers and practitioners from various disciplines and application areas will help the different disciplines relevant to data linkage to coalesce.

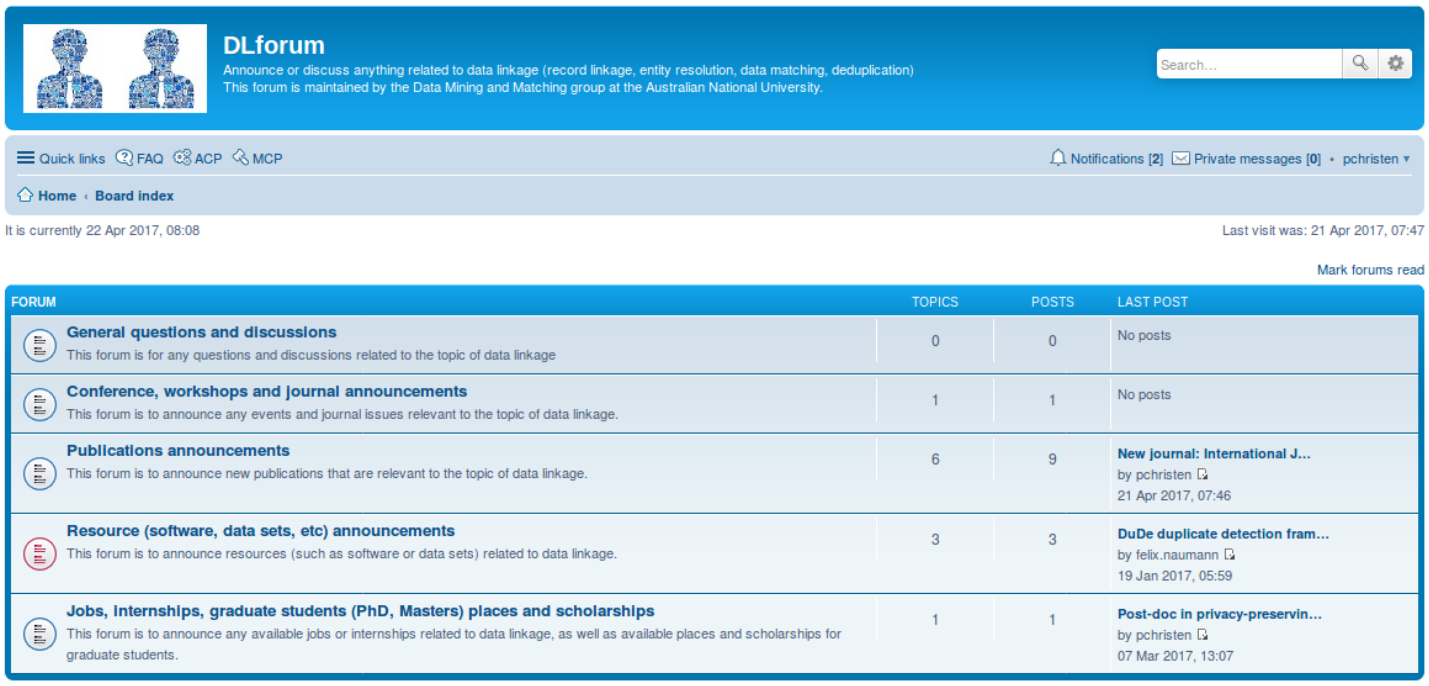
Online forums, which are also known as message, bulletin or discussion boards, have been popular since the early days of the Internet. Today, online forums are used very widely, ranging from companies that use such forums for customer support (for examples see: https://www.phpbb.com/showcase), universities that employ them to support online learning [14] as well as counselling [15], and communities of interest that use them to encourage communication among their members (see for example: https://www.kaggle.com/discussion). Pendry and Salvatore [16] have recently studied the use of online forums and found that engagement in such forums can lead to benefits for both individuals and a community, and also lead to improved off-line engagement because it allows users to better identify with other users as well as the community that is served by an online forum. Based on these findings, our hope is for the *DLforum* to be able to fill the current gap of online resources for the multi-disciplinary data linkage community and facilitate improved communication and knowledge dissemination within and across disciplines. We describe the *DLforum* in detail in the following section.

Methods

As described above, an online discussion forum needs to allow users to access and post content in an efficient and easy to use way. Before starting the development process we investigated the applicability of different Internet forum software systems. In our investigations we considered several features such as open source availability, programming language, licensing, continued development, and the availability of extensions (plugins that facilitate for example improved security or statistical evaluation of a forum's usage). Based on these features we selected the widely used open source discussion forum software *phpBB* (freely available from: http://www.phpbb.com), where *phpBB* stands for *PHP Bulletin Board*. PHP is a popular programming language especially suited for Web development.

We selected *phpBB* because of several reasons. This software facilitates the easy development of online discussion forums, including the management of users and messages posted. *phpBB* provides hierarchical sub-forums with a flat message structure which allows users to initiate new topics and respond to previous posts by other users. *phpBB* also provides extensive user management functionalities which allow administrators to implement various security features and techniques to avoid automatic registrations by spammers or malicious users. The software also allows administrators to control and monitor postings by users by providing different authorisation mechanisms. These measures improve the overall security of the forum and the privacy of the legitimate registered users. *phpBB* has been developed since the early 2000s and is one of the most widely used free bulletin board software available. It is used by thousands of Websites and millions of users on a daily basis. The development of *phpBB* is an on-going effort where the software is continuously updated and improved by a community of developers.

Figure 1: Main page of the DLforum showing the five main forum areas



The current structure of the *DLforum* is based on discussions with data linkage practitioners and researchers, including discussions at the workshop "Data Linkage: Techniques, Challenges, and Applications" in Cambridge in September 2016, and by inspecting online forums in related areas such as the Kaggle discussion forums (see: https://www.kaggle.com/discussion). Kaggle is a data science competition Website, where datasets provided by companies, governments, and not-for-profit organisations are being analysed by volunteers in a competitive manner. Kaggle provides a system to conduct data science competitions, and has helped to connect organisations with specific problems and tasks with a large community of students and volunteer data scientists who are keen to solve such real-world data science problems. The winners of such competitions, those that for example are able to develop predictive machine learning algorithms with the highest accuracy for a given task, are being awarded financial rewards or job opportunities. The Kaggle Website contains a highly active forum where competitions and general data science topics are intensively discussed. The number of topics and discussions on Kaggle is over 30,000. With this inspiration, we developed the DLforum to serve the data linkage community.

Figure 1 shows the current structure of the *DLforum*, which was set to contain five high-level topic areas (discussion areas), as described in detail below. Depending upon future demand by the data linkage community, changes to these areas as well as additional main level topic areas can easily be incorporated into the *DLforum*.

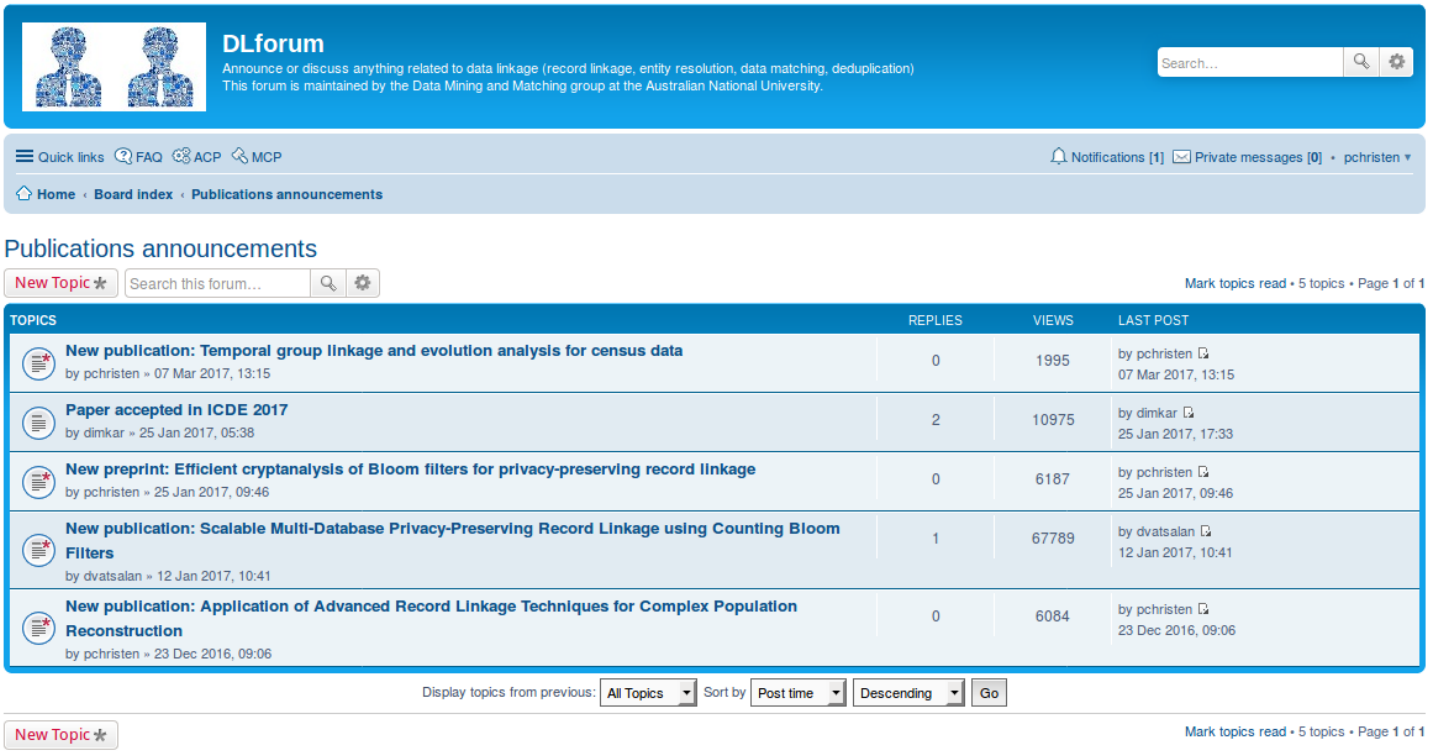
1. **General questions and discussions:** In this area, users are invited to post any questions related to data linkage, ranging from technical and methodological to practical questions. Users are encouraged to provide answers to any questions, and we hope that over time there will be certain discussion threads that lead to collaborations across disciplines to tackle common problems and challenges in data linkage.

2. **Conference, workshop and journal announcements:** Here we encourage the organisers of any event relevant to data linkage to announce their events (such as calls for paper submissions to workshops, conferences or journals, or the actual programs of events). Over time this will hopefully lead to a one-stop site where researchers and practitioners can find venues that are relevant to data linkage. Special journal issues relevant to data linkage are also highly encouraged to be announced in this forum.

3. **Publication announcements:** Publications relevant to any aspect of data linkage can be announced in this area, with links provided to the actual publication to facilitate access to the work by the data linkage community. The nature of the discussion forum also allows users to post questions and comments on publications, such as querying details of a publication or enquiring if a dataset used in a publication is available to other researchers. Figure 2 shows a list of such publication announcements.

4. **Resource announcements:** An important aspect of a practical research area such as data linkage is the sharing of resources such as test datasets and software. It is currently the case that data linkage researchers and practitioners might have datasets they need to have linked but they lack the technical expertise or software to link these datasets (as might be the case for health or social science researchers). On the other hand, researchers in statistics and computer science might have developed

Figure 2: Publication announcements forum area showing several individual posts.



novel data linkage algorithms and/or software but they commonly do not have access to real datasets to test them on. This discussion board aims to connect these often disparate groups.

5. **Jobs, internships, graduate student places and scholarships:** This final area is for announcements of any opportunities for employment and student support, but also for students who are seeking a job placement in data linkage.

As illustrated in Figure 2, each of these topic areas brings the visitor to a Web page which shows the posts (announcements and discussions) in the given area. Users can sort these posts in different order, search for content of interest, and select posts to view their actual details, as can be seen in Figure 3. If they have registered with the *DLforum* (as discussed below), they can reply to posts, and also send private messages to the user who has written a certain post.

The *phpBB* software provides users with various settings with regard to how much information about them is made visible on the forum, who can contact them, and so on, thus providing privacy and confidentiality to users. *phpBB* also has a variety of settings available that allow registered users to customise how they see messages and if they want to receive notifications when new messages are posted in a certain forum, or replies are added to one of their own posts.

While accessing and reading any of the messages in the *DLforum* is open to anyone, users who are interested in posting messages to the *DLforum* are required to register. This is free and required to eliminate the possible abuse of the *DLforum* by individuals and automatic systems that aim to spam public forums with abusive or otherwise non-relevant content. A request for signing up to the *DLforum* will be vetted by one of the moderators of the forum, and if required an email confirmation will be sought from a new user. However, registering with DLforum and any posting of messages on the *DLforum* is free. Posts will be inspected on a regular basis by the moderators, and any inappropriate post will be deleted and the corresponding author reminded of the appropriate use of the forum.

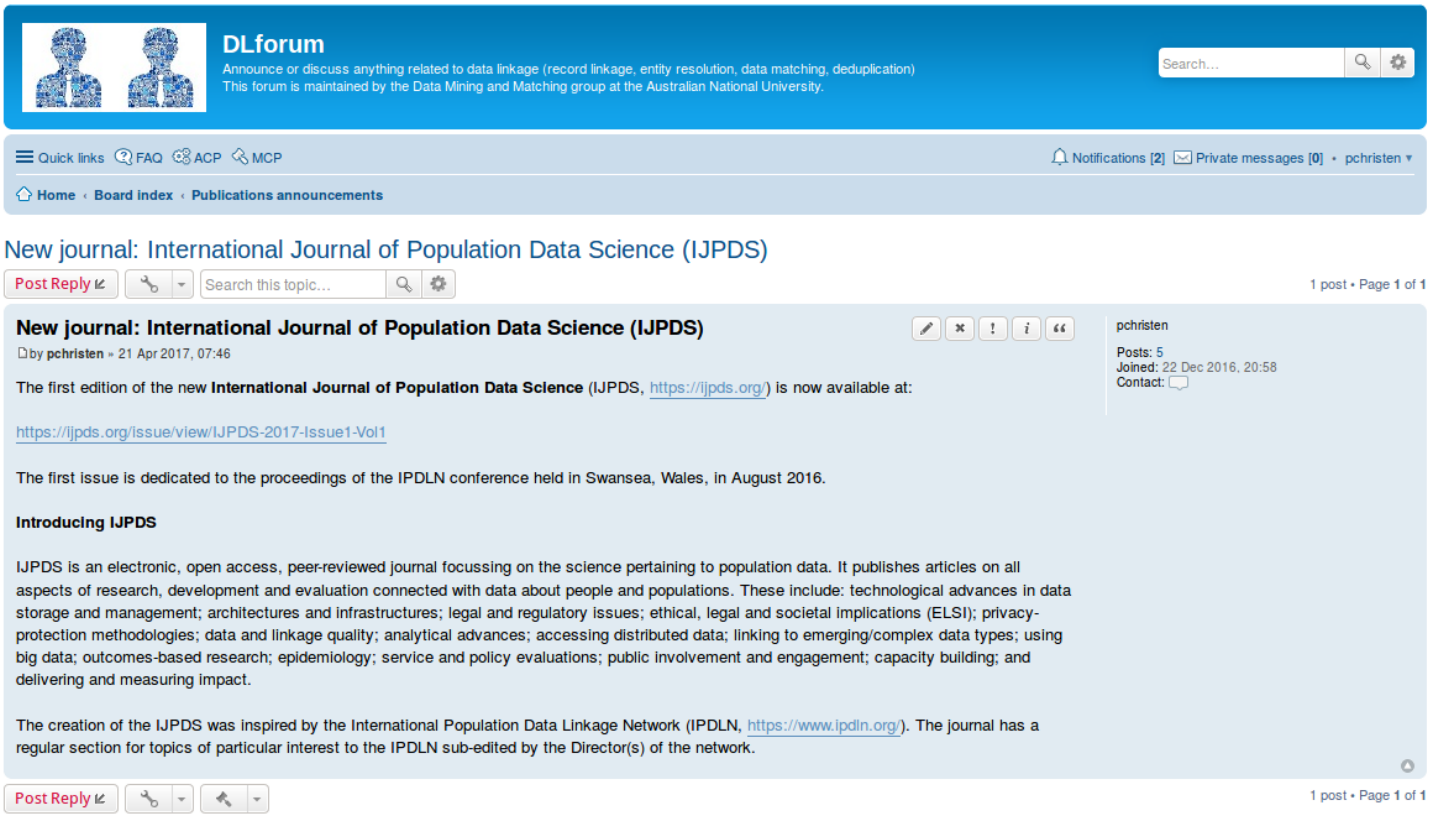
Currently, the *DLforum* is monitored by several moderators from the Australian National University who approve new users and assess the content published in the system. If needed additional moderators will be sought from the data linkage community. The *DLforum* has received approval from the appropriate IT Services and management at the Australia National University, where the *DLforum* is currently hosted.

We invite the interested reader to explore the *DLforum* at: https://dmm.anu.edu.au/DLforum/

Discussion and Conclusion

In this article we have presented the *DLforum*, a free online forum (and bulletin board) that we hope will provide a multidisciplinary platform for data linkage researchers and practitioners to interact with each other and importantly across disciplines. Our aim with the *DLforum* is to provide a single unifying website for all matters concerning data linkage, to allow the different data linkage related communities to coalesce, for individuals to learn about advances and events in other areas of data linkage, and find open positions in industry,

Figure 3: An individual example post announcing the IJPDS.



government or academia, and for students to find interesting challenges as well as support for their open research questions. We believe the *DLforum* addresses the current gap of online resources for the multi-disciplinary data linkage community and helps to improve communication and knowledge dissemination and thereby create an online community of practice.

So far, the *DLforum* has only been advertised to a small number of test users, and therefore it only contains a small number of posts. We hope with the publication of this IJPDS article there will be a larger uptake of the *DLforum* by data linkage practitioners and researchers. We plan to analyse and evaluate the usage of the forum as the number of registered users, as well as the number of topics posted and replies to posts, increases over time. We also plan to survey the *DLforum* user-base at some stage in the future (once we have reached a critical number of users) to gather feedback on how the *DLforum* can be improved (with additional high level topic areas, new functionalities and features, and so on). We aim to further promote the *DLforum* to the data linkage community in future conferences and workshops.

We encourage all readers of the IJPDS with an interest in data linkage to sign up to the *DLforum*, to post any relevant news to the forum, to participate in discussions on the forum, and to advertise the forum to their colleagues and students who might be interested in data linkage. Our hope is that the *DLforum* over time will become the main online resource for data linkage researchers and practitioners to disseminate, exchange, and obtain information about advances in this exciting and crucial topic in population data science.

Acknowledgements

Initial ideas of the *DLforum* were discussed at the Isaac Newton Institute (INI) for Mathematical Sciences at the University of Cambridge. The authors would like to thank the INI for support and hospitality during the "Data Linkage and Anonymisation" (https://www.newton.ac.uk/event/dla) programme (EPSRC grant EP/K032208/1). Peter Christen was also supported by a grant from the Simons Foundation. This work was also partially funded by the Australian Research Council (DP130101801).

References

1. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Berlin: Springer; Aug 2012.

2. Harron K, Goldstein H, Dibben C. Methodological developments in data linkage. London: John Wiley & Sons; Sep 2015.

3. Herzog TN, Scheuren FJ, Winkler WE. Data quality and record linkage techniques. New York: Springer; 2007 May.

4. Lawrence S, Giles CL, Bollacker K. Digital libraries and autonomous citation indexing. IEEE Computer. Jun 1999;32(6):67-71. DOI: 10.1109/2.769447

5. Rahm E. Discovering product counterfeits in online shops: A big data integration challenge. ACM Journal of Data and Information Quality. Sep 2014;5(1-2):3. DOI: 0.1145/2629605

6. Dunn HL. Record linkage. American Journal of Public Health and the Nations Health. Dec 1946;36(12);1412-6.

7. Newcombe H, Kennedy J, Axford S, James A. Automatic linkage of vital records. Science. Oct 1959;130(3381);954–9. DOI: 10.1126/science.130.3381.954

8. Newcombe H and Kennedy J. Record linkage: Making maximum use of the discriminating power of identifying information. Communications of the ACM. Nov 1962;5(11);563–6. DOI: 10.1145/368996.369026

9. Fellegi I and Sunter AB. A theory for record linkage. Journal of the American Statistical Association. Dec 1969;64(328);1183–1210.

10. Dong, XL and Srivastava, D. Big Data integration. Synthesis Lectures on Data Management. San Rafael; Morgan and Claypool; 2015. DOI: 10.2200/S00578ED1V01Y201404DTM040

11. Kum, HC, Krishnamurthy A, Machanavajjhala A, Ahalt, SC. Social genome: Putting big data to work for population informatics. IEEE Computer. Jan 2014;47(1);56-63. DOI: 10.1109/MC.2013.405

12. Hall R and Fienberg S, Privacy-preserving record linkage. In: Domingo-Ferrer J and Magkos E editors. Privacy in Statistical Databases – PSD 2010; Sep 22-24; Corfu, Greece. Springer LNCS, p. 269-283.

13. Vatsalan D, Sehili Z, Christen P, Rahm E. Privacy-preserving record linkage for Big Data: Current approaches and research challenges. In Zomaya AY and Sakr S, editors. Handbook of Big Data Technologies. Cham, Switzerland: Springer; Feb 2017. DOI: 10.1007/978-3-319-49340-4_25

14. Kaur M. Using online forums in language learning and education. Inquiries Journal. 2011;3(03).

15. Richards D. Features and benefits of online counselling: Trinity College online mental health community. British Journal of Guidance and Counselling. Aug 2009;37(3):231-42. DOI: 10.1080/03069880902956975

16. Pendry LF, Salvatore J. Individual and social benefits of online discussion forums. Computers in Human Behavior. Sep 2015;50:211-20. DOI: 10.1016/j.chb.2015.03.067