Harvard Data Science Review • Issue 1.2, Fall 2019

Data Linkage: The Big Picture

Peter Christen[1]

[1]**Australian National University, Canberra, Australia**

The MIT Press

Published on: Nov 30, 2020

DOI: https://doi.org/10.1162/99608f92.84deb5c4

In the past, a data scientist might have expected to work on a single, well-defined, cleaned, and curated static data set. Today, however, data scientists increasingly gain interesting, novel, valuable, and unexpected knowledge through combining multiple data sets (Dong & Srivastava, 2015). These data sets might come from diverse sources: an organization could generate them internally, or their generation could be external, perhaps from statistical agencies, commercial partners, or government departments. Data scientists can even gather data from websites or social networking platforms. Data sets like these often have uncontrollable quality, including errors, missing values, and selection distortions. Moreover, such data could be outdated, or, if dynamic, updated on a regular or irregular basis. Some data sets are highly sensitive or confidential, and thus have restricted access. Cleaning and pre-processing such data sets, as well as integrating them to answer questions, pose formidable challenges. Such challenges are collectively known as *data wrangling*.

This short article covers integrating diverse data sets, with a specific focus on *how to identify and link records that correspond to the same entity* within one or across several data sets. This task—which goes by names such as *data linkage*, *record linkage*, *data matching*, or *entity resolution* (Christen, 2012; Herzog, Scheuren & Winkler, 2007)—is a crucial step of data integration. The other two steps are *schema matching* or *mapping* (identifying which attributes/fields across database tables contain the same type of information), and *data fusion* (merging a set of linked records that correspond to the same entity into a single coherent, complete, and up-to-date record that represents that entity). *Duplicate detection,* or *deduplication*, is when records about the same entity need to be identified in a single data set (such as in longitudinal data sets) (Naumann & Weiss, 2010).

A major challenge when linking records from diverse sources is the lack of a common *entity identifier* across the data sets. For example, patient identifiers might not occur in all the databases to be linked. As a result, data scientists can employ so-called quasi-identifiers (QIDs) to identify and link records about the same entity. For databases that contain records about people, applicable QIDs include names, addresses, phone numbers, dates of birth, and so on. Applications such as online bibliographic databases and digital libraries use the titles, author names, and venues of publications as QIDs (Christen, 2012).

Data linkage applies to many domains. Traditionally, national statistical agencies have used such techniques to link census records of persons or households over time. In the health system, data linkage has consolidated patient records from diverse health care providers, or collected over time (for example, to identify all records of births from the same mother). More recent applications include the reconstruction of (historical) populations based on linking birth, marriage, and death certificates (Bloothooft et al., 2015). Accomplishing this for a large population over a significant period of time permits new types of studies showing how education, health, and migration policies have influenced a society, and how genetic effects influence the health of people over several generations. Data linkage techniques are also of crucial importance in crime and fraud detection, as well as national security. Here, data integrated from diverse sources can identify patterns of suspicious behavior and the individuals who follow those patterns. Other examples include linking all records by the same

author in bibliographic databases so as to allow funding agencies to better understand the impact of researchers, and linking consumer products and their descriptions in online comparison-shopping sites (where different online shops often have somewhat different names and descriptions of the same consumer product) to allow customers see all offers for the same product across different online shops.

Halbert Dunn first described the idea of data linkage in 1946 as a *book of life* for each individual (Dunn, 1946), and since the 1950s, several research domains have developed data linkage techniques using QIDs (Newcombe, Kennedy, Axford, & James, 1959). Initial ad-hoc techniques developed by statisticians and health researchers compared pairs of records and calculated numerical *matching weights* for different attributes (or fields) based on the importance of these fields. The seminal work by Fellegi and Sunter (1969) on *probabilistic data linkage* provided a sound theoretical basis. They developed an optimal decision approach to classifying record pairs into *matches* (pairs assumed to refer to the same entity), *non-matches* (pairs assumed to refer to different entities), and *potential matches* (where no clear match or non-match decision could be made), based on the similarity of their QIDs' values. These potential matches then go out for *clerical review*, where domain experts manually assess if a record pair is a match or a non-match.

Fellegi and Sunter's (1969) idea is based on how much information can be learned about a pair of records by comparing their QIDs. For example, matching people solely on their 'gender' attribute is less likely to identify the same person in each database than is matching them on their 'surname' attribute. Furthermore, in data sets from the US, UK, or Australia, for example, two records with surname values `Smith' are less likely to refer to the same person than two records with surname value `Dijkstra', because more people in these populations have the common surname `Smith' than the rare surname `Dijkstra'. Fellegi and Sunter showed that calculating and combining such matching weights for all attributes in a set of QIDs will allow an optimal linkage decision, given certain assumptions and limitations.

This basic probabilistic data linkage approach has been refined in various ways. One example is the use of approximate string-matching techniques (Christen, 2012; Herzog, Scheuren & Winkler, 2007) which can calculate approximate similarities between somewhat similar strings (e.g., `Gail' and `Gayle,' or `Kristina' and `Christine'). Another is tackling scalability to large databases by *blocking* (Christen, 2012). This involves grouping records according to some criterion and only comparing records that lie in the same group or block (e.g., compare only those records with the same zip- or postcode value).

In more recent times, computer science researchers have approached data linkage from a traditional *binary supervised classification* perspective (with the two classes being matches and non-matches, but with no potential matches), or from a *clustering* perspective. In the former, training data are required in the form of known matching and non-matching record pairs, along with the similarities between their QID values. Note, however, that data linkage is generally a highly unbalanced classification problem because (even after blocking has been applied) there will be many more non-matching than matching record pairs. For example, assuming two data sets of 1,000 records each, where each record refers to one entity, there will be a maximum of 1,000

matching but a minimum of 999,000 non-matching record pairs if no blocking has been applied. This means that simple misclassification rate is not a suitable measure for linkage quality. Instead, measures based on *precision* and *recall* are often used (Hand & Christen, 2018).

One challenge with both traditional probabilistic data linkage and modern supervised pairwise classification techniques is that they can lead to inconsistent results due to nontransitivity (Christen, 2012). If we have classified record pair (R1, R2) as a match and (R1, R3) as a match, then the pair (R2, R3) should also be a match. However, when the pairs are looked at separately, the third pair might have been classified as a non-match – or even not compared at all if blocking was used. This issue can be overcome by considering data linkage classification as a clustering problem (Hassanzadeh, Chiang, Lee, & Miller, 2009) where the aim is to group all records that refer to the same entity into one cluster. Alternatively, it can be tackled by so-called *collective entity resolution* techniques (Bhattacharya & Getoor, 2007) that consider relationships between records in the classification step (such as common co-authors for a certain author, or shared addresses for people living in the same household) and seek an overall optimal linkage solution. While such clustering and collective techniques have shown in certain cases to lead to high quality linkage results, their computational requirements currently make these techniques impractical for linking very large data collections.

In many application areas the most common type of records to be linked describe people (e.g., patients, customers, tax payers), where their QIDs might include confidential personal details such as names, gender, addresses, and dates of birth. Protecting the privacy of such confidential data during the linkage process is the aim of the emerging research area of *privacy-preserving record linkage* (Hall & Fienberg, 2010). In this, the parties involved in a linkage learn only limited information about which record pairs are classified as a match, but nothing about the actual records and the values from any other party involved in the linkage. Privacy-preserving data linkage involves a combination of encoding and encryption techniques, for example using Bloom filters (Schnell, Bachteler, & Reiher, 2009) and balancing the challenges of approximate matching and scalability against the requirement for adequate privacy protection and provable security (Vatsalan et al., 2017). Privacy-preserving data linkage applications are now starting to be employed in several countries worldwide, mainly in the health sector to link sensitive health data collections (Boyd, Randall & Ferrante, 2015).

Data linkage is far from a completely solved problem, and the modern big data era is presenting various new challenges (Christen, 2012; Dong & Srivastava, 2015). In many applications, data sets are not static, but records about entities are updated or added continuously (e.g., patient records in hospitals, or social security records in a government database). This requires novel linkage techniques that can deal with dynamic data. Certain applications require query records to be linked to a large database of entity records in (near) real time (such as police officers on the street trying to identify a suspect in a database of known criminals). Modifications in how matching weights are calculated are required when temporal data are being linked, since each record will have a time-stamp attached showing when it was generated or last updated. For example, if it is known that a high proportion of a population changes address over time, then a different address may not

mean it is less likely that two otherwise matching records refer to different individuals. Increasingly, there are also applications that require the linking of multiple – potentially into the hundreds or even thousands – of data sets. They may also have different schemas, formats, and types. Combined with the increasing demand for protecting the privacy of personal and otherwise sensitive or confidential data (Vatsalan et al., 2017), there is a rich field of open research challenges in data linkage to be tackled.

Disclosure Statement

Peter Christen has no financial or non-financial disclosures to share for this article.

References

Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 5. https://doi.org/10.1145/1217299.1217304

Bloothooft, G., Christen, P., Mandemakers, K., & Schraagen, M. (Eds.). (2015). *Population reconstruction*. Springer. https://doi.org/10.1007/978-3-319-19884-2

Boyd, J. H., Randall, S. M., & Ferrante, A. M. (2015). Application of privacy-preserving techniques in operational record linkage centres. In *Medical Data Privacy Handbook* (pp. 267–287). Springer. https://doi.org/10.1007/978-3-319-23633-9_11

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. *Data-Centric Systems and Applications*, Springer. https://doi.org/10.1007/978-3-642-31164-2

Dong, X. L., & Srivastava, D. (2015). Big data integration. *Synthesis Lectures on Data Management*, 7(1), 1–198. Morgan and Claypool. https://doi.org/10.2200/S00578ED1V01Y201404DTM040

Dunn, H. L. (1946). Record linkage. *American Journal of Public Health and the Nations Health*, *36*(12), 1412–1416.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Hand, D. and Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, *28*(3), 539–547. https://doi.org/10.1007/s11222-017-9746-6

Hall, R., & Fienberg, S. E. (2010). Privacy-preserving record linkage. In *Proceedings of the 2010 international conference on Privacy in statistical databases* (pp. 269–283). ACM.

Hassanzadeh, O., Chiang, F., Lee, H. C., & Miller, R. J. (2009). Framework for evaluating clustering algorithms in duplicate detection. *Proceedings of the VLDB Endowment*, *2*(1), 1282–1293. https://doi.org/10.14778/1687627.1687771

Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer. https://doi.org/10.1007/0-387-69505-2

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management*, *2*(1), 1–87. Morgan and Claypool. https://doi.org/10.2200/S00262ED1V01Y201003DTM003

Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, *130*(3381), 954–959. https://doi.org/10.1126/science.130.3381.954

Schnell, R., Bachteler, T., & Reiher, J. (2009). Privacy-preserving record linkage using Bloom filters. *BMC Medical Informatics and Decision Making*, *9*(1), 41. http://doi.org/10.1186/1472-6947-9-41

Vatsalan, D., Sehili, Z., Christen, P., & Rahm, E. (2017). Privacy-preserving record linkage for big data: Current approaches and research challenges. In *Handbook of Big Data Technologies* (pp. 851–895). Springer. https://doi.org/10.1007/978-3-319-49340-4_25
