Knowledge of the Law in the Big Data Age
G. Peruginelli and S. Faro (Eds.)

doi:10.3233/FAIA190028

Dealing with Privacy Issues in Data Integration: Scenarios for Official Statistics

Piero Demetrio FALORSI [a], Brunero LISEO [b] and Monica SCANNAPIECO [a]
[a] *ISTAT - Istituto Nazionale di Statistica (Italy)*
[b] *Sapienza Università di Roma (Italy)*

Abstract. The increase of data availability poses new challenges and suggests new interesting road to public and private data producers and providers. The European Commission acknowledged these opportunities that can significantly boost European competitiveness in the global market and in scientific research. One of the cornerstones of the process to build a common European data space is the possibility to access and share public and publicly funded data. This task has many important different goals: 1) citizens' secure access to and sharing of health data; 2) improving and innovating healthcare solutions based on mobile applications; 3) multiple uses of public sector information; 4) sharing scientific information, in order to facilitate the dissemination of results across countries; 5) Economics: Business to Business (B2B) data sharing, which considers the availability of "non personal machine-generated data". The new challenges suggest new problems to be faced both on a legislative and on a methodological ground. From a legal perspective, data exchange between public Institutions and private agents requires a detailed national legislative framework, still missing in many European countries. From a methodological perspective, the interaction between public and private data holders poses complex problems: 1) privacy preserving record linkage: how to guarantee that the linkage of personal data coming from different sources will not jeopardize the privacy of single citizens and/or companies; 2) the use of linked data as input to more sophisticated statistical analyses without unplanned information disclosure. Secure Multiparty Computation (SMC) techniques can play a role in this respect. In this Chapter we describe how the Italian National Institute of Statistics (Istat) is facing the new challenges and what are the most important steps to take in the next future.

Keywords. multiple sources, data protection, record linkage

1. Background: Official Statistics and National Statistical System

The role of official statistical national agencies has become more and more important in the last decades, especially in the most developed countries. The computational and methodological advances in the ability of managing, exchanging and combining different statistical information sources have dramatically increased the importance of such agencies which are going to play a central role in the process of stimulating data use and evidence based decision making by the governments.

The term 'national statistical system' (NSS) refers to a country's producers of official statistics, generally a national statistical institute (NSI) and other institutions and administrations producing official statistics. National statistical system and, in particular, national statistics institutes (NSIs) have several duties.

In Italy, the Italian National Institute of Statistics (Istat) commitments are regulated by the so called National Statistical Program – also known as Sistan[1], which has a 3 years agenda – established with the Legislative Decree No. 322/1989, and it represents the legislative tool for planning statistical activities to be carried out by Sistan bodies, including Istat, in order to satisfy international standards and country's information needs.

The National Statistical Program determines, for example, what are the statistical outputs which must be publicly available, to what level of detail the above statistical outputs need to be disseminated, in terms of variables, classification rule and aggregation level, in such a way to not cause disclosure risk and, at the same time, to guarantee a satisfactory information level. The National Statistical Program is divided into two volumes and an annex:

- The first volume is devoted to describe the way in which statistical information evolves and changes (framework information; information gaps by sector; aggregated costs of NSP).
- The second volume is much related to privacy issues and provide information and limitations about the activities processing personal data: it contains guidelines on personal, sensitive and/or judicial data processing.
- The annex is very technical: it establishes the methodologies which determine what level of disaggregation can be used for each single variable to be disseminated.

One of most relevant issues in the production of statistical information is the possibility of using administrative information *in lieu* of ad hoc statistical surveys. The reasons for doing that are multiple: administrative lists usually provide larger sample sizes, very reduced costs and minimal response burden.

On the other hand, they also imply a series of disadvantages, including measurement issues, and the fact that administrative data are generally not collected after an a-priori statistical design, their quality may be very low. In particular, they could not exactly answer the questions that a National Statistical Institute may want to ask.

A relatively new trend is to link both data sources, administrative records and survey data, to enhance the level of information and to open the way to more sophisticated data analyses.

It is not rare, however, that administrative data are collected by other public and/or private agencies and the operation of data exchange and linkage must be precisely regulated, especially at micro-level, where disclosure issues may arise.

Nowadays, in Italy, Istat can acquire administrative micro-data of public ownership for its mission. The entire data treasure consists, in 2018, of 478 administrative archives, owned by 94 different subjects.

The other subjects of Sistan may exchange micro-data from other subjects of the system which are strictly necessary for their specific mission.

Also, Sistan subjects are not allowed to release micro-data which are produced or acquired by others. There are a specific regulation and protocols for the supply of micro-

[1] www.sistan.it.

data for research. Even more complex are the relations between private agencies and Sistan members; in these cases it is necessary to build specific partnerships, in agreement with the data protection Authority.

2. Record Linkage Techniques and Connections with Privacy-preserving Issues

Today, the need for increasingly detailed and timely statistical information is shared by several international or supranational (European Union, European Central Bank, International Monetary Fund, etc.) and national (National Statistical Service, Ministries, Regions, etc.) bodies, and private users too. In this respect, increasing computational potential provides important opportunities. It is now possible to collect and maintain massive amounts of statistical data obtained from the integration of survey data and administrative information.

In this context, a significant problem is represented by the need to merge various data archives, possibly in view of the fulfillment of different goals. The awareness of the scientific community of this problem is testified by the numerous international symposia on "combining data from different sources". The main statistical approaches employed to address these problems can be classified as:

- Record linkage.
- Statistical matching.

The latter technique seeks to derive integrated statistical information by combining information from different datasets, in which only some variables are observed twice, and no overlapping of observed units is necessary. In this Chapter, we will only focus on the former approach. Record linkage refers to the use of specific algorithms that aim to identify pairs of records, corresponding to a single statistical unit, that are present in different databases. The same problem is addressed – albeit in a more general manner – in information technology literature, as the problem of integrating non-aggregated databases. In this context, relevant issues are (i) the construction of a general framework (ii) the detection and specification of semantic relationships between non- homogeneous data sources (iii) the characterization of data quality factors and (iv) the reconciliation of datasets from different sources, in order to construct a representation that is coherent with the relevant general framework and quality requirements. The following applications of record linkage methodologies deserve mention:

1. The construction and maintenance of a list of statistical units, to be used as a 'reference population' in sample or total surveys. In this context, it is important to identify units that feature in more than one database.
2. The merging of two or more databases to obtain a single archive, which is more informative at a non-aggregated level. This makes it possible to perform statistical analyses that would be otherwise impossible.
3. The use of several data sources for the improvement of the overall survey's 'coverage'. The methodological implications of these problems are not yet well-developed, but it is certain that the information provided by administrative data archives can be of great assistance in this regard.
4. Population size estimation problems via capture-recapture methods. A relevant example is the estimation of the under-coverage given by a complete census: this is usually performed via a linkage analysis between total survey data and an ad hoc post-enumeration survey [1].

5. The evaluation of the validity of a disclosure method, to protect access to administrative data from the risk of identification of single units by an intruder [2]; [3].

The use of record linkage techniques poses several interesting problems, in both methodological and computational terms. From the methodological point of view, the very definition of a statistical model (the description of how comparisons between records are performed) is still debated: see, for example, [4]; [5]; [6]; [7]; [8]. From a computational perspective, problems become formidable once the databases reach a large size (over 100 units). In these cases, comparisons are performed only between records that have the same values for certain 'blocking variables', which are assumed to have been recorded without errors. A broadly satisfactory solution of these problems appears, therefore, crucial.

In recent years, we have experienced a great proliferation of new Bayesian methodologies and, especially, an increasing number of statistical applications performed from a Bayesian perspective. The main reason for this trend lies in the development of Markov chain Monte Carlo (MCMC) methods which enable building and calibrating virtually any statistical model, regardless of complexity. This opportunity has made Bayesian methods much more appealing and visible in many areas of application, including official statistics. NSIs have several important and complex tasks; for their practical implementation, different kinds of – more or less – subjective operational decisions must be taken. For example, several important economic and social indexes are the result of procedures that at least implicitly involve the use of complex statistical models. Nevertheless, the result of a statistical analysis performed by NSIs 'must' be objective or, at least, should be perceived as such by users.

Bayesian concepts can be important for official statistics when (i) important prior (or extra- experimental) information on the variables of interest exists, and cannot be exploited adequately in a classical inference framework; and (ii) even when prior information is missing, a Bayesian analysis can be required, because a classical approach cannot provide answers unless strong assumptions, not easily tested, are introduced. In these situations, a Bayesian analysis enables at least a sensitivity analysis, to quantify the influence of the assumptions on the inferences made.

In general, from the point of view of statistical methodology, merging two (or more) data files can be important for two reasons:

- Per se, to obtain a larger and integrated reference dataset.
- To enable performance of a subsequent statistical analysis, based on the additional information obtained, that cannot be extracted from either of the two individual data files.

As already noticed, record linkage and disclosure techniques are intimately related because potential identification of units from record linkage techniques may disclose – even accidentally – sensible information at micro-level.

Consider a case of international crime investigation, where different databases from different countries and agencies are compared and linked to gain information. Similar examples arise in biomedical science where the use of integrated data may help in the detection of adverse drug reactions [9].

The disclosure and sharing of databases containing sensitive information is a very complex task and it must be regulated both from a legal and from a methodological perspective. Following [10], we define the privacy-preserving record linkage in the following way.

"*There are k different owners of databases D(1), ... , D(k). Each owner aims at determining which units in his/her databases match some units of other databases according to a decision criterion which compares strings belonging to different databases*".

Each owner does not wish to reveal his/her own actual records with any other party. They are willing to share with other parties, the actual values of some selected attributes of the record pairs which are classified as matches by the decision rule.

This can be done in several different ways. A first important distinction is between techniques which involve a third party or not. In a three party protocol, a third institution is involved in performing the linkage and it represents a filter between the two data owner. In a two-parties scenario, the two owners directly interact and sophisticated techniques are needed in order to avoid the disclosure of sensitive information during the linkage process.

3. Integration Scenarios: How to Preserve Privacy When combining Multiple Sources

In this Section we will highlight different scenarios where techniques like privacy-preserving record linkage could find application.

We envision four scenarios that can support the 'generic' information sharing need, namely: (i) private set intersection (PSI); (ii) private set intersection with enrichment (PSI-E); (iii) private set intersection with analytics (PSI-A); (iv) private data mining.

More specifically:

- Private Set Intersection (PSI): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an exact join to A and B without revealing any unnecessary information about their individual databases. That is, ideally, the only information learned by P1 about B is A∩B and vice versa (Figure 1).
- Private Set Intersection with Enrichment (PSI-E): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an exact or approximate join to A and B without revealing any unnecessary information about their individual databases. After that, they wish to enrich joined records with variables by both parties. At the end of the process P1 will learn additional P2 variables on A∩B and vice versa (Figure 2).
- Private Set Intersection with Analytics (PSI-A): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an analytics function to the intersection of A and B in a private way. At the end of the process, the only information learned by the parties (beyond the keys of the records belonging to the intersection) is the result of the analytics function (Figure 3).
- Private data mining (PDM): Let P1 and P2 be parties owning (large) private databases A and B. The parties wish to apply an analytics function to the union of A and B without revealing any unnecessary information about their individual databases. At the end of the process, the only information learned by the parties is the result of the analytics function (Figure 4).

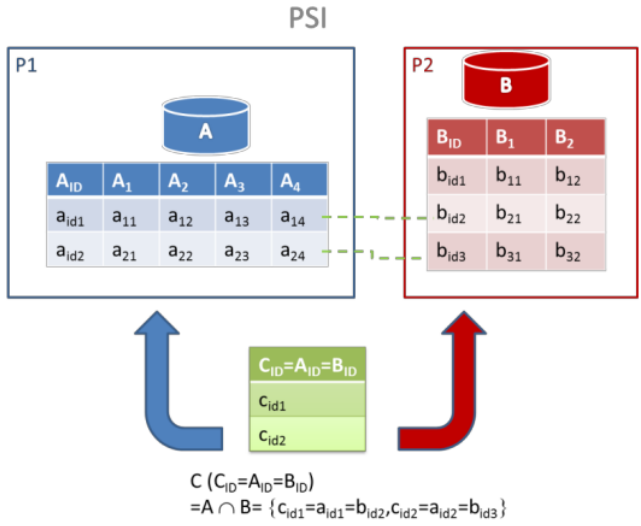

Figure 1. Example of PSI, P1 learnes that a_{id1} and a_{id2} are also owned by P2 and P2 learns that b_{id2} and b_{id3} are also owned by P1

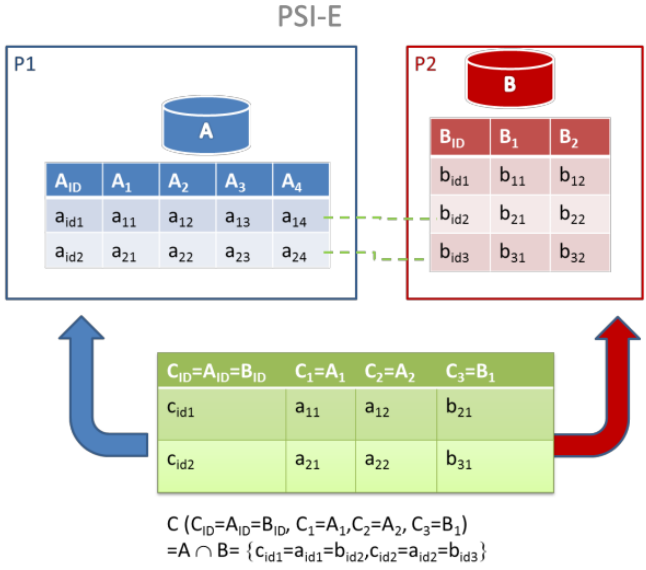

Figure 2. Example of PSI-E, P1 learnes that a_{id1} and a_{id2} are also owned by P2 and also their values for attribute B_1. Similarly, P2 learns that b_{id2} and b_{id3} are also owned by P1 and also their values for attributes A_1 and A_2.

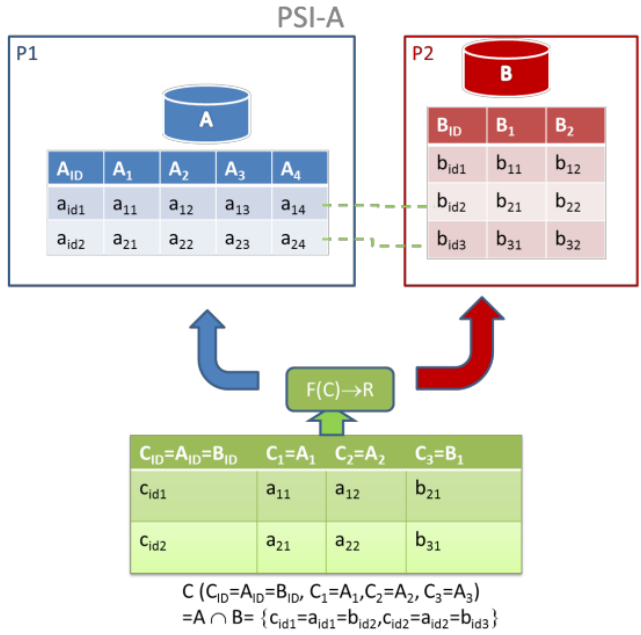

Figure 3. Example of PSI-A, P1 learnes the result R of the analytics function F applied to the intersection C and P2 gets the same

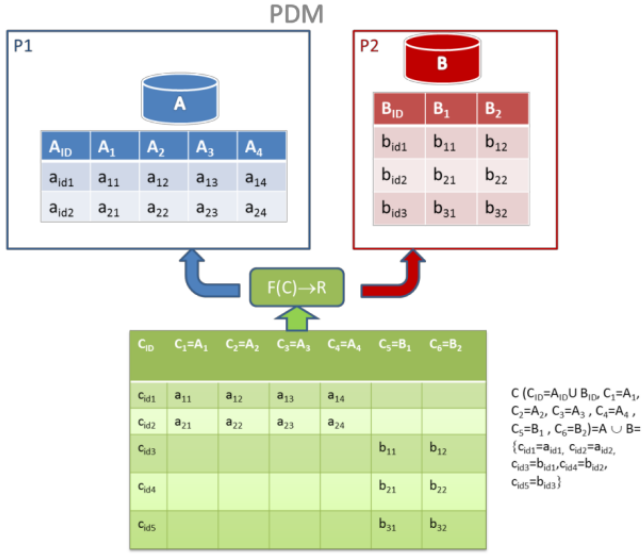

Figure 4. Example of PDM, P1 learnes the result R of the analytics function F applied to the union C and P2 gets the same

4. Conclusions

Even if Official Statistics already has a defined regulatory framework for privacy protection, the enforcement of privacy preserving measures at technical level in integration scenarios is necessary. In this Chapter, we have illustrated some specific scenarios for integrating data in a privacy-preserving way that could exploit techniques like privacy preserving record linkage. In order to implement solutions for these scenarios several aspects should be considered, namely: organizational, regulatory, methodological and technological. On the basis of our experience, one key factor is to have multidisciplinary teams working together on the specific objective. The investment from statistical organizations should be carefully considered and planned. However, there are many drivers that push for such investments, the main one being the fact that an organization can have access to data owned by another organization without directly accessing it, and without violating any privacy constraints. This fosters the flexibility of statistical organizations in answering statistical users' needs, while at the same time saving money and reducing response burden.

References

[1] Winkler, W. E. (1986). *Record Linkage of Business Lists*. Energy Information Administration, U.S. Dept. of Energy. Technical Report.

[2] Duncan, G. & Lambert, D. (1989). The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, *7*(2), 207-217.

[3] Winkler, W. E. (1998). Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, *Research in Official Statistics*, 1, 87-104.

[4] Fellegi, I. P. & Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

[5] Copas, J. B. & Hilton, F. J. (1990). Record linkage: Statistical Models for Matching Computer Records. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 153*(3), 287-312.

[6] Belin, T. R. & Rubin, D. B. (1995). A Method for Calibrating False-match Rates in Record Linkage. *Journal of the American Statistical Association, 90*(430), 694-707.

[7] Fortini, M., Liseo, B., Nuccitelli, A. & Scanu, M. (2001). On Bayesian Record Linkage. *Research in Official Statistics, 4*(1), 185-198.

[8] Tancredi, A. & Liseo, B. (2011). A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems. *The Annals of Applied Statistics*, *5*(2B), 1553-1585.

[9] Elmagarmid, A. K., Ipeirotis, P. G. & Verykios, V. S. (2007). Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *19*(1), 1-16.

[10] Vatsalan, D., Christen, P., O'Keefe, C. M. & Verykios, V. S. (2014). An Evaluation Framework for Privacy-preserving Record Linkage. *Journal of Privacy and Confidentiality*, *6*(1), 35-75.