Exploring Solutions for Linking Big Data in Official Statistics

Tiziana Tuoto, Daniela Fusco and Loredana Di Consiglio

Abstract Official statistics has acknowledged the value of big data and has started exploring the use of diverse sources in several domains. Sometimes, big data objects can be easily connected to statistical units. If a unit identifier is available, the opportunity to link big data to existing statistical micro data can allow enlarging the content, the coverage, the accuracy and the timeliness of official statistics, for example Internet-scraped data could be used with this aim. In this setting, new challenges arise in data integration with respect to linking administrative data. In this work, we describe a real case of integration of web scraped data and a statistical register of agritourisms specifying the novelties and challenges of the procedure.

Keywords Big data · Internet-scraped data · Data integration
Data linkage · Farm register

1 Introduction

In the recent years, Official Statistics has acknowledged the value of Big data and has started exploring the use of diverse sources in several domains. For some of these external sources, the object can be easily associated with a statistical unit of the target population. In those cases, if a unit identifier is available and shared with the NSI, the opportunity to link Big data to already existing statistical data at micro-level can allow enlarging the content, the coverage, the accuracy and the timeliness of official statistics. In the case of the Internet-scraped data, there is a

T. Tuoto (✉) · D. Fusco · L. Di Consiglio
Istat, Via Cesare Balbo 16, Rome, Italy
e-mail: tuoto@istat.it

D. Fusco
e-mail: dafusco@istat.it

L. Di Consiglio
e-mail: diconsig@istat.it

© Springer International Publishing AG, part of Springer Nature 2018

C. Perna et al. (eds.), *Studies in Theoretical and Applied Statistics*,
Springer Proceedings in Mathematics & Statistics 227,
https://doi.org/10.1007/978-3-319-73906-9_5

great potential to enrich the information of the traditional statistics. In this setting, new challenges arise for data integration experts in official statistics, due to the deep differences with respect to the familiar framework in which administrative data have been integrated for a long time in order to produce statistical outputs, e.g. statistical business and population registers. In this study, exploiting a real case as a starting point, we describe novelties and challenges in integrating Internet-scraped data with traditional statistical datasets, from the entity extraction and recognition phases to the unit matching algorithms. As case study, we propose the linkage of Internet-scraped information with data related to agritourisms, as reported in the statistical Italian farm register, that is obtained by the integration of several administrative sources. In order to overcome some limits of the so far well-established linkage procedures, we propose to explore new techniques not yet introduced in the official statistics production system. Finally, we devote the due attention to the output quality evaluation, to better analyse benefits and risks of the integration and to allow the analysts to take into account potential integration errors in subsequent analyses.

2 Standard Solutions in Official Statistics

The combined usage of data coming from different sources is a current practice in official statistics since some decades. Administrative data is largely used for building up statistical registers, as benchmark at aggregated level, as auxiliary information and for replacement of survey variables [4]. When data are available at micro level, record linkage techniques are widely applied in order to recognize in different sources the same real world entity even in absence of a unique identifier.

The most widespread statistical approach to record linkage is based on the Fellegi and Sunter [5] theory and several statistical institutes have designed and developed tools to be able to face record linkage problems arising in linking survey and administrative data. An up-dated critical review of methods and software for record linkage is in Tuoto et al. [15] where the proliferation of methodologies and tools in recent years is interpreted according to assigned criteria, namely flexibility of the tools with respect to the support to input/output formats, extensibility, maturity, supported language and coverage of functionalities related to identified sub-phases in which it is possible to organize a record linkage process to reduce its recognized complexity.

Since 2006, the Italian National Institute for Statistics (Istat) has designed and developed RELAIS, an open source toolkit providing a set of techniques for facing record linkage projects [14]. As principal feature, RELAIS is designed and developed to allow the combination of different techniques for each of the record linkage phases, so that it can provide solutions for different linkage problems on the basis of application and data specific requirements.

In ten years, the RELAIS toolkit has been applied in several situations, i.e. when dealing with:

- data of different quality coming from administrative sources (as resident permits, hospital admissions, road traffic accidents, etc.);
- huge amount of data (i.e. the 60 million of people involved in the Italian population census at the time of the post-enumeration survey)

Previous experiences with data related to agriculture have already shown that these data have some peculiarities with respect to the record linkage task, essentially related to the inaccuracy of the "name" of the statistical units and the difficulties in identifying detailed address in rural areas. In fact, in the agricultural field the "name" could be the farm name (like the business name) or the farm operator name. This is particularly the case when such information come from administrative sources and refer to units that only indirectly are associated to the target statistical units, for instance because the sources were compiled for different purposes, e.g. fiscal and benefit purposes. Moreover, when extended large families operate farms, different family members could appear as reference for tax, benefits or as land owners. Furthermore farm addresses in rural areas are often the generic name of the land and are often shared for all the units and houses located, increasing the difficulties in linking sources and resulting in imperfect matching.

3 Issues in Linking Big Data

The definition of a linkage strategy requires selecting the set of the most discriminant common variables. Linkage methods typically compare different variables of the entities using a set of distance measures. The resulting similarity scores may be combined using different aggregation functions. If the data sources use different variable value representation formats, values have to be normalized by applying transformations prior to the comparison. Designing a good linkage strategy is a non-trivial problem as the linkage expert needs to have detailed knowledge about the data sources in order to choose appropriate variables, data transformations, distance measures together with good thresholds as well as aggregation functions.

Currently, a very large number of information is available from the web. Official statistics has acknowledged the value of Internet-scraped data and has started exploring their use in several domains (for instance in statistics on ICT use in enterprises [1] and tourism [6, 11]). In most cases, data coming from the web is not directly comparable with data collected and organized by National Statistical Institutes and a lot of work is needed to create integrated data.

In fact, data can be scraped directly from the website where several information at unit level can be harvested (company, agritourism in this case). On the other hand, this model requires first identifying the websites and then it has to face with different queries and different formats obtained from each website. Otherwise, data can be scraped from "hub", website hosting and describing a plurality of units (for instance, hotel, agritourism, etc.), in this case the number of information that can be

achieved is smaller than in the previous case and conditional on the available information on the hub.

In addition, it is likely that the Internet-scraped data are not standardized or codified. In some sense, official statistics are already prepared to face this kind of problems but the pre-processing phase is still a very time-consuming process and a lot of work is needed to identify models that can easily support the data reconciliation, the management of the complexity and to allow the data integration step. Other typical problems can be related to the low quality of data and to changes in the data model of the external source.

So, in order to integrate web scraped data in the official statistics production process, a system is required for data ingestion and reconciliation that allows managing a big data volume of data coming from a variety of sources. The statistical production system needs to produce the ontology and the big data architecture, and the mechanisms for the data verification, reconciliation and validation.

The main issue of this task seems to be the definition of the reconciliation algorithms and their comparative assessment and selection. A reconciliation step is needed to deal with the variety and variability of data, e.g. with the presence of several different formats, with the scarce (or non-existing) interoperability among semantics of the fields of the several datasets. Generally, in order to reduce the ingestion and integration cost, by optimizing services and exploiting integrated information at the needed quality level, a better interoperability and integration among systems is required [2, 8, 12].

This is desirable but not always possible, due to lack of communication, agreement, operability between actors. This problem can be partially solved by using specific reconciliation processes to make these data interoperable with other ingested and harvested data. On the other hand, this approach does not solve the problem since instances can be not interoperable and linked together. For example, a street names coming from two different sources physically identifying the same street may be written in different manner creating a sematic miss-link. These problems have to be solved as well with reconciliation processes. On the other hand, the web data sources may report a "commercial" name as "name" for identifying unit, something appealing for web searchers and potential customers, while in the official data the unit is reported with different name unrelated with the previous one.

The whole linkage activity includes processes of data analysis for ontology modeling, data mining, formal verification of inconsistencies and incompleteness to perform data reconciliation and integration. Among the several issues, the most critical aspects are related to the ontology construction that enables deduction and reasoning, and on the verification and validation of the obtained model.

The reconciliation phase may find advantages in relying on a repository or dictionary, for example of existing Street in a Town, but this kind of products is expensive to build up or to acquire. In fact, a relevant process of data improvement for semantic interoperability is related to the application of reconciliations among the entities associated for instance with locations as streets, civic numbers and localities. Unfortunately, it is not always possible to perform reconciliation at street

number level, i.e. connecting an instance that uniquely identifies a street number on a road, the finest level of localization; sometimes, the reconciliation is only at street-level, with less precision, or at the worst at municipality level. On this regard, there are different types of inconsistencies for which reconciliation did not return results, both for the lack of references into the scraped data (some streets and civic numbers can be missing or incomplete) and for lack of correspondences into the repository/dictionary. Finally, it could happen to have location entities which result as wrong and not reconcilable due to (i) the presence of wrong values for streets and/or locations, and (ii) the lack of a consistent reference location.

To summarize, the reconciliation phase requires different steps, for instance starting from searching for correspondences at the finest level (i.e. the street number), but allowing for correspondences at higher level of disaggregation (i.e. street, municipality) in the further steps. Finally, applying a manual correction and cleaning or manual search of non-identified matches into a list of probable candidates suggested by the previous disregarded results.

During the reconciliation step, there may be cases where no connection among the data are caused by a different encoding of the instances, for instance the name of the municipality. To support the reconciliation process each time new data are available, they should be automatically completed with the correct Istat municipality code.

Another very common issue in integration is related to the existence of multiple ways to express the toponymy qualifiers, (e.g. Piazza and P.zza) or parts of the proper name of the street (such as Santa, or S. or S or S.ta): this issue can be overcome thanks to support tables, inside which the possible change of notation for each individual case identified are inserted.

At the end of this huge and complex reconciliation process, it is possible to define the most appropriate linkage strategy, i.e.:

- Choose appropriate variables.
- Specify and tune the parameters: the distance measures together with good thresholds.
- Define the proper model for aggregating scores functions.

However, definitively, the effectiveness of the linkage process is dramatically affected by the output quality of the pre-processing phase.

An alternative to standard linkage methods (Fellegi and Sunter, [5, 9]) is represented by supervised learning algorithm which employs genetic programming in order to learn linkage rules from a set of existing reference links. For instances, [10] proposed GenLink, which is capable of matching entities between heterogeneous data sets which adhere to different schema. By employing an expressive linkage rule representation, the algorithm learns rules which:

- Select discriminative properties for comparison.
- Apply chains of data transformations to normalize property values prior to comparison.

- Apply multiple distance measures combined with appropriate distance thresholds.
- Aggregate the result of multiple comparisons using linear as well as non-linear aggregation functions.

Following genetic programming, the GenLink algorithm starts with an initial population of candidate solutions which is iteratively evolved by applying a set of genetic operators. The basic idea of GenLink is to evolve the population by using a set of specialized crossover operators. The applicability and efficacy of this kind of solution in official statistics context has to be studied and tested.

Finally, whatever will be the linkage strategy, for a proper usage in official statistics context, the verification and validation of results is a key aspect, as well as the provision of the quality indicators of process and products, in order to allow other data analysts to perform the correct analyses on the integrated data. This task is often very expensive in terms of time and resources.

4 A Case Study: Linking Internet-Scraped Data of Farmhouse to the Farm Register

As case study we propose an integration of Internet-scraped data regarding agritourism with data reported in the Farm Register built up by Istat. The final aim of this integration is the use of Internet information for statistical purpose, in particular to update and integrate some agricultural data collected in the Farm Register.

In fact, in order to produce harmonized and comparable statistics, one of the Eurostat core recommendations is to define and set-up a Statistical Farm Register (SFR). SFR represents a key element for the Agricultural Statistical System and the basis for sample selection. In Italy, it is built by integrating several administrative sources (Integrated Administration and Control System, Animal register, Tax declaration on agricultural land, land cadaster, Chamber of Commerce, Tax on Value Added on agricultural income) and some statistical sources (Business Register, Agricultural Census, Agritourism survey, Quality product survey).

Big data could be used to update the SFR, permitting the production and the periodical dissemination of statistics related to the activities and to the services offered by the agritourism farms, at a minimum cost. The choice of agritourism topic depends on the presence of portals in the web, an important perspective in this experimental phase.

So, the initial and most important target of web scraping is represented by the different websites acting as "hubs" (hosting and describing a plurality of Agritourism), in general maintained by Regions, or by private organizations, and containing information regarding name, address, geographic coordinates, telephone, e-mail, prices, offered services, etc. A specific scraping application is developed for each hub; this permits to collect all the semi-structured information so to compare it

to the official data set. At the end, three hubs are considered, among the most popular websites providing this kind of information.

As specified in the previous section, this unrefined information requires a laborious pre-processing activity because they are unusable for integration in the way they are scraped. Then it is necessary to standardize the variables for each hub dataset. In the internet-scraped data the several information are not delimited by separators or fixed length of the fields, so it is necessary to recognize address, province, town, region, country distinguishing those by the agritourism name. It is necessary to recognize the different variables and then codify municipalities with the Istat code. Addresses are validated using the following software http://www.egon.com/en/solutions/address-validation.html

In some cases it is not possible to identify a correct address and the observations without a normalized address have been deleted (about 35%).

Furthermore, to increase the significance of the company names in each file the most common words (i.e. Agriturismo, Azienda Agricola, Affittacamere, ect.) have been eliminated.

In addition, it is necessary to make a first linkage process between the three datasets from the three different hubs in order to obtain a single deduplicated dataset comparable with the Farm Register.

The linkage activities have been performed using RELAIS. Several linkage strategies have been applied, the most effective in revealing matches without introducing false matches is based on the following step:

1. Data cleaning—preparation of the input files (pre-processing) as above described; as well known in official statistics the preparation of input files is the first phase and requires 75% of the whole effort to implement a record linkage procedure, in this case the pre-processing step is particularly huge and expensive, requiring almost the 95% of the whole time.
2. Choice of the common identifying attributes (matching variables); after the previous phase, it is important to choose matching variables that are as suitable as possible for the considered linking process. We use: company name, address, province, municipality and region.
3. Search space creation/reduction; to reduce the complexity it is necessary to reduce the number of pairs to compare. We choose Blocked Simhash function [13, 3], which combines the Blocking Union method (using region variable) with SimHash (using company name/address variable).
4. Choice of decision model; in absence of a unique identifier we decide for a probabilistic model according to the Fellegi and Sunter [5] theory.
5. Choice of comparison functions; Comparison functions measure the “similarity” between two fields. We use equality for province and municipality variables and, to overcome non identical strings, the Inclusion3Grams for address and company name variables. This function have been chosen because it takes into account the number of 3-length grams in common between the two strings, and the target is the 3-grams amount of the shortest string.

To obtain the best result, the linkage strategy has been designed in two iterations: at the beginning we used the company name and region for the search space creation with Blocked Simhash function. So only the company name has been used as match variable with province and municipality.

In the second process, on the set of unlinked records of the first step, the address has been used instead of the company name, with the same model, functions and thresholds.

The chosen strategy allowed to link 2765 units, 37.8% of the smaller file, represented by the 7301 farms from the web. The farms in the Farm register are 13503.

The result is still adequate because there are some aspects to be taken into account. The frame of SFR Agritourism is 13000 units on a total of 20000 existing for the Agricultural Ministry. The difference may be caused by the failure of the address normalization we have described. The SFR was built in 2013, while the websites are updated frequently and the number of agritourism in the portal is likely to be increasing. On the other hand, agritourism obtained by web scraping the portals might be false farms and for this reason not included in the SFR.

The double linkage process has allowed to overcome two types of difficulties. Addresses referred to the headquarter in a source and to the farm manager residence in the other one could be resolved. Similarly, the issue of different names in the two sources was overcome.

The comparison function Inclusion 3 Grams was very functional to this issue. In fact, often in the company name string was also the name of the farm manager in one file. With other similarity functions it could easily result in a link failure. This function was also very useful for addresses, especially addresses in German (province of Bolzano) written in different ways in the two files.

5 Concluding Remarks and Future Work Direction

The first attempt to link web-scraped and traditional data in official statistics allows underlining potentials and issues connected to this operation.

The first evidence highlights the role played by the pre-processing phase and the data cleaning/reconciliation activity. Traditionally these procedures require a lot of work, 75% of the whole effort to implement a record linkage procedure, according to [7] but in this case the time and work devoted to this task were even more than ¾ of the whole effort. Ignoring this task, however, may compromise the effectiveness of the following steps. In fact, an attempt on raw not-preprocessed data resulted in no matches identified.

The achieved match rate may seem low but the knowledge of data can explain these results. Moreover, the comprehension of the reasons why the match rate is low in this application may provide the main leverage to improve it. The main issues in this specific field are:

- Very often the farm names are different in the sources. Generally in internet we can find the name of the agritourism, instead in administrative or statistical sources the farm is registered with the company name. So, even with an accurate data cleaning, it is impossible to identify the same unit if the address is not accurate and does not provide enough linking power;
- However, very often the farm addresses are inaccurate. They are main roads, shared by large lands, like *contrada*, *regione*, etc., generally without a street number, so it is not possible to identify an exact and unique position.
- In some sources the farm headquarter address is recorded, while in other ones we find the farm manager residence address. When the residence of the farm manager does not correspond to the farm headquarter address, it is not possible to improve the matching results even after address normalization.

It is quite clear that some of the previous issues are not specific of the Internet-scraped data but they could generally arise with data coming from any kind of sources not designed for statistical purposes. Moreover, in this specific field the Internet data itself may provide a strong improvement to the linkage procedure by means of the easy availability of further information on the geographical references of the interest units. The use of geoinformation in linkage procedure seems a promising direction to explore.

Finally, it seems that this kind of linkage activity needs first of all to define an incremental process able to analyze, integrate and validate each added data sources. The main point to facilitate this task is related to the availability of mechanism for automatizing the data interpretation, verification, reconciliation and coding. In the Big data context, whit several sources different with respect to many aspects, this step cannot be managed with traditional solution, requiring a hard human resource involvements. In parallel, techniques for data linkage, different with respect to the traditional ones, may allow the comparison of less structured data. At the end, the statistical validation of the linkage results and the measurement of output quality need to be assessed.

References

1. Barcaroli, G., Scannapieco, M., Nurra, A., Scarnò, M., Salamone, S., Summa, D.: Internet as Data Source in the Istat survey on ICT in Enterprises. Austrian J. Stat. **44**, 31–43 (2015)
2. Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R.: OWLIM: a family of scalable semantic repositories. Semant. Web J. **2**(1) (2011)
3. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proceedings ACM STOC 2, pp. 380–388. Montreal, Quebec, Canada (2002)
4. Citro, C.:(2014) From multiple modes for surveys to multiple data sources for estimates. Surv. Method.
5. Fellegi, I.P., Sunter A.B.: A theory for record linkage. J. Am. Stat. Soc. **64** (1969)
6. Fuchs, M., Höpken, W., Lexhagen, M.: Big data analytics for knowledge generation in tourism destinations—a case from Sweden. J. Destination Mark. Manage. (2014)

7. Gill, L.: Methods for Automatic Record Matching and Linking and their Use in National Statistics. National Statistics Methodological Series No. 25. London: Office for National Statistics. http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9224 (2001)
8. Gupta, S., Szekely, P., Knoblock, C., Goel, A., Taheriyan, M., Muslea, M.: Karma: a system for mapping structured sources into the semantic web. In: Proceedings of the 9th Extended Semantic Web Conference (ESWC2012)
9. Jaro, M.: Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa. Fla J. Am. Stat. Soc. **84**(406), 414–420 (1989)
10. Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. Web semantics: science, services and agents on the WorldWideWeb **23**, 2–15 (2013)
11. Heerschap, N., Ortega, S., Priem, A., Offermans, M.: Innovation of tourism statistics through the use of new big data sources. Technical Paper, Statistics Netherlands (2014)
12. Hepp, M.: GoodRelations: an ontology for describing products and services offers on the web. In: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008), Acitrezza, Italy. vol. 5268. Springer LNCS, 29 Sept–3 Oct 2008, pp. 332–347
13. RELAIS 3.0 User Guide. http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/relais (2015)
14. Tuoto T., Cibella N., Fortini M., Scannapieco M., Tosco L., (2007) RELAIS: Don't Get Lost in a Record Linkage Project, Proc. of the Federal Committee on Statistical Methodologies (FCSM: Research Conference. Arlington, VA, USA (2007)
15. Tuoto, T., Gould, P., Seyb, A., Cibella, N., Scannapieco, N., Scanu, M.: Data Linking: A Common Project for Official Statistics in Proceedings of Conference of European Statistics Stakeholders Rome 24, 25 Nov 2014