VERSITA

Journal of Official Statistics, Vol. 29, No. 1, 2013, pp. 165–169, DOI: 10.2478/jos-2013-0010

Discussion

Frauke Kreuter[1]

This special issue on "Systems and Architectures for High-Quality Statistics Production" is a stimulating resource for statistical agencies and private sector data collectors in a challenging time characterized by massive amounts of data, from a variety of sources, available in varying intervals, and with varying quality.

Traditionally, statistical products were created from a single source, most often through surveys or administrative data. However, neither surveys nor administrative data alone can match the data needs of today's society. In addition, the need to reduce the costs of data production necessitates that multiple sources are used in combination. The need to reduce costs also necessitates the streamlining of production cycles, and the increasing difficulties in data collection itself require such systems to be much more flexible than they have been in the past. Increasingly, these reasons are driving statistical agencies and private data collectors to redesign their entire data production cycle. The examples in this special issue from Statistics Netherlands and Statistics New Zealand demonstrate such developments in government agencies; the example from RTI reflects efforts visible among private sector data collectors. This commentary will highlight some issues of general interest related to organizational challenges, and some that create the basis for reproducible research and are therefore of general interest to the research community.

1. Organizational Challenges

A common phenomenon among many national statistical agencies and also many private data collectors is an *organizational structure around projects or topics*, rather than around tasks. Such an organizational structure is often referred to as one involving stovepipes or silos, and all three articles comment on the presence of such structures in their organizations. In each of these organizations, the *silo structure* resulted in a multiplicity of software solutions, duplication of work, a lack of dissemination of knowledge from silo to silo on particular tasks, and a lack of standardization in survey products and documentation.

To overcome the disadvantages of silos, each organization restructured the statistics production process. At RTI, the restructuring revolved around a core of software tools that in the future will be used across all survey data collection projects as well as a common data storage system. Similar approaches were taken by Statistics New Zealand. Statistics Netherlands went a step further and also changed the structure of the organization itself,

[1] Joint Programme in Survey Methodology, University of Maryland, Ludwig Maximilian University, Munich and Institute for Employment Research (IAB), Email: fkreuter@survey.umd.edu

implementing modular production lines, and moving away from parallel production lines. For Statistics Netherlands, the modular approach allowed for better use of scarce resources, such as expertise in specialized fields, like sampling. Moving away from silos also greatly facilitated the use of the same metadata (data about data) and documentation across systems, streamlining training needs and ultimately reducing costs.

All three organizations report on the difficulties they encountered in implementing the changes to existing data production systems. Most of these challenges are not unique in this context and can be found elsewhere. One of these challenges is the *adoption of a new system during ongoing production*. More subtle, but nevertheless important problems reported in these articles are those arising from the *lack of participatory decision making*. Because individual units are most familiar with specific processes, input from these units is essential to developing new routines. The Statistics Netherlands article points out that communication across people working in Information Technology and survey production is crucial to a successful operation. Another common challenge is the continuous updating of the data within the systems and the availability of data from the system to each production unit at appropriate times.

All three articles, but in particular the two from the national statistical agencies, emphasize the growing demand for *the use of administrative data* on its own or in combination with survey data. However, the descriptions in the articles could spend more time detailing on how such combinations are achieved. When data from multiple sources are combined, individual records do not always have unique and identical sets of identifiers. In such cases, probabilistic record linkage is used. Privacy concerns often result in identifiers being encrypted, creating another challenge for unified systems. Here, new methods are being developed to allow for privacy-preserving record linkage (see e.g., Schnell et al. 2009). In some cases, clearing houses (within or across agencies) are used to allow for record linkage with unencrypted identifiers. It would be interesting to know how the various agencies handle this problem.

When administrative data is used for statistic or data production, its *quality* is of paramount importance. The three articles here lack some clarity on how the quality of administrative data is assessed and ensured beyond edit checks. It is not unlikely that several data sources will have overlapping sets of variables and that those variables are of different quality, depending on their source and the importance of a given variable when the data are entered. Incorporating the perspective of efficient data linkage into the building of statistics production systems is important.

2. Reusability and Reproducibility

One of the key motivations for the restructuring of the statistical production process is the possibility to reuse existing knowledge, data and code across statistical products and across organizational silos. The focus of all three organizations on employing unified software systems and comparable documentation systems not only enhances reproducibility but creates the basis for quality assurance. For example, at Statistics New Zealand, undocumented manual repairs to the data threatened the quality of the final product in the past. The article comments on new software systems that allow automatization of routine tasks (to avoid hand-coding) and the automated creation of data

tables for storage and reports. Such automatization is often missing, and even organizations with smaller projects would benefit from unified software code stored for long-term use and reproducibility.

Metadata play an important role in this process. In order to move data across silos or production processes in an automated fashion, certain standardizations must be in place. One example of this kind of standardization is the introduction of common labels for corresponding variables. As simple as this might seem from the outside, each article commented on the difficulty of achieving such standardization. With long-standing data products in particular, such changes can be tedious for individual silos. Having achieved a common set of metadata that allows easy exchange of data within the organizations is therefore a major accomplishment.

However, this agenda could be pushed even further. Thinking of statistical data users located outside the statistical agencies, standardization across statistical offices would also be advantageous. Outside users are much more likely to use data from various sources, maybe even various countries, than users within the same agency. Outside users often attempt to combine or compare such data. Large resources are then needed (over and over again) when researchers make data comparable *post hoc*. The standards developed by the Data Documentation Initiative (DDI 2012) could help overcome this problem. DDI has developed standards for describing data from the social, behavioral, and economic sciences that support the entire research data life cycle, including the collection, processing, distribution, and archiving of data. For example, a set of controlled vocabularies sets metadata standards in semantics (which elements should be defined), and content (what values should be assigned to elements and how). A naïve user of data might think, for example, that all surveys ask questions about race and ethnicity in the same way, whereas quite the opposite is true even within a single organization. To date, primarily academic entities have subscribed to the DDI standards, but some Research Data Centers (RDC) affiliated with government organizations, such as the RDC at Statistics Canada or the RDC affiliated with the German Institute for Employment Research (IAB), are in the process of designing their metadata systems according to such international standards (Bender et al. 2011).

Of course, if analysts frequently shared their data preparation and analysis code, efficiency gains in such repetitive work could be accomplished even without the international standardization of metadata. Fortunately, some journals already encourage, if not require, the dissemination of analysis code. However, the inability to search journals' supporting online material make the dissemination of analysis code through free code repositories such as GitHub (http://github.com) and SourceForge (http://sourceforge.net) much more desirable (for a larger discussion, see Peng 2011). Publicizing the code used by statistical agencies to transform raw data into data products can not only increase efficiency, but also contributes to quality assurance by making the process reproducible.

In addition to metadata and shared analysis code, paradata – data created as a by-product of the production process – are another key element in quality assurance, and one that has to be considered in the construction of unified systems. Many agencies are interested in using survey production-based paradata to inform decisions about cases during ongoing fieldwork and to allow for a timely shift of cases from one production mode to another. One example – beyond the ones discussed here – is the U.S. Census Bureau, which is currently creating a flexible and integrated system to use paradata in

ongoing survey management (Thieme and Miller 2012). However, paradata can also be used to monitor the processes within the agencies themselves. Paradata examples of the latter could be records on edits done to a particular case or data set (already prominently used at RTI), production times for a particular module (e.g., in the Dutch case), or the number of data manipulations necessary to move raw data into long-term storage in flexible data bases (e.g., in the case of New Zealand). In this sense, the whole statistical production process can be monitored through paradata. The development of useful paradata is currently still in its infancy and their measurement error properties are unknown (see also Kreuter 2013).

3. Outlook

It is interesting that each organization developed its own system, and did not seem to collaborate with others in the design process. This special issue is a valuable step towards fostering such collaborations. These articles will help disseminate information about the new system developments and hopefully spur dialogue and interaction amongst the various data collection entities. My wish would be for organizations to share some of the system infrastructure when possible. While it would probably be difficult to share the entire IT infrastructure, code pieces could very well be provided and would be of use to others – for example, code pieces necessary to extract paradata from data collection software, or algorithms used in call scheduling or models used for responsive designs. A larger discussion would also benefit from developments in other fields or those that develop across disciplines. To mention just one example, the British company Jisc aims to create a Virtual Research Environment (http:jisc.ac.uk/whatwedo/programmes/vre.aspx).

The discussion about the system development also highlights anew the need for people trained with a total survey error perspective. The total survey error perspective (TSEP) deconstructs error in a survey estimate into its constituent pieces, such as nonresponse error and errors due to measurement. Being aware of the various sources of error and their likely contributions to a given estimate based on previous research will allow for informed trade-off decisions as data are combined within the system. A framework similar to the TSEP could also be used to understand the errors in administrative data, although much more work is needed in this area. One obvious place to look to begin to create an error taxonomy for administrative data is the study of questionnaire design, for administrative data are often also collected through self-administered forms that resemble surveys.

Survey methodologists can also help identify the effects of changes made to one production module on other modules. This will be important, for example, when unified systems are used to move cases in a flexible way, mid-field, from one mode of data collection to another. In addition, survey methodologists can use their tools to assess the possibilities of merging various data sources with different error structures (a point not mentioned in the three articles), and their expertise in the analysis of such linked data (Lahiri and Larsen 2005). Finally, survey methodologists can contribute their knowledge about issues of data linkage that need to be addressed at the data collection stage, for example, when asking for linkage requests (see e.g., Calderwood and Lessof 2009; Sakshaug et al. 2013).

4. References

Bender, S., Dieterich, I., Hartmann, B., and Singula, D. (2011). FDZ-Jahresbericht 2009/2010. Available at: http://doku.iab.de/fdz/reporte/2011/MR_06-11.pdf (accessed February 14, 2013).

Calderwood, L. and Lessof, C. (2009). Enhancing Longitudinal Surveys by Linking to Administrative Data. Methodology of Longitudinal Surveys, P. Lynn (ed.). New York: Wiley.

DDI (2012). What is DDI? Available at: http://www.ddialliance.org/what (accessed February 14, 2013).

Kreuter, F. (ed.) (2013). Improving Surveys with Paradata: Analytic Use of Process Information. New York: Wiley.

Lahiri, P. and Larsen, M. (2005). Regression Analysis with Linked Data. Journal of the American Statistical Association, 100, 222–230.

Peng, R. (2011). Reproducible Research in Computational Science. Science, 334, 1226.

Schnell, R., Bachteler, T., and Reiher, J. (2009). Privacy-Preserving Record Linkage Using Bloom Filters. BMC Medical Informatics and Decision Making, 9(41).

Sakshaug, J., Tutz, V., and Kreuter, F. (2013). Placement, Wording, and Interviewers: Identifying Correlates of Consent to Link Survey and Administrative Data. Survey Research Methods (forthcoming).

Thieme, M. and Miller, P. (2012). The Center for Adaptive Design. Presentation to the National Advisory Committee on Racial, Ethnic, and Other Populations. October 25, 2012, Washington, D.C.