SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES AND OFFICIAL STATISTICS

Rory Fitzgerald[1] and Sarah Butt[2]

[1] *City University London, London, UK; r.fitzgerald@city.ac.uk*
[2] *City University London, London, UK; sarah.butt.1@city.ac.uk*

Abstract
Synergies for Europe's Infrastructures in the Social Sciences (SERISS) is an EC funded project to strengthen and harmonise the collection and curation of social science data across Europe. The project brings together leading European Research Infrastructures in the social sciences – ESS ERIC, SHARE ERIC and CESSDA AS – alongside GGP, EVS and the WageIndicator Survey. Together, these infrastructures are undertaking a programme of work to address key challenges in cross-national data collection (including representativeness and translation), break down barriers between infrastructures through the creation of shared online tools, and equip social research for the future by exploring new forms of data collection.

This paper highlights some of the key areas where outputs from SERISS could feed into the production of high quality official statistics and the realization of the European Statistical System Vision 2020. The paper also suggests areas for future collaboration between SERISS infrastructures, national statistical institutes and other producers of official statistics.

Keywords: cross-national surveys; data quality; collaboration

1. Introduction

A robust socio-economic evidence base, drawing on official statistics, wider social survey data on individuals' attitudes and behaviour and new and emerging data sources, is essential if Europe is to be able to tackle grand societal challenges such as climate change, immigration, an ageing society, fertility changes and youth unemployment amongst others. Synergies for Europe's Infrastructures in the Social Sciences (SERISS) is an EC funded project which brings together Europe's leading research infrastructures in the social sciences in order to strengthen and harmonise data collection and curation and ensure that the social sciences can continue to contribute effectively to this evidence base.

The intended aims and outputs of the SERISS project align closely with the European Statistical System (ESS) Vision 2020 which aims to ensure that official statistics are "fit for the future" and that the ESS is guided by quality, promotes efficiency through collaboration and embraces new opportunities arising from the digital revolution, as well as engaging with users of statistics (ESS, 2014).

In this paper we provide an overview of the SERISS project and explain how the different strands of work being undertaken are relevant to the five key areas included in ESS Vision 2020, particularly those associated with data quality, efficient processes, and new forms of data. The paper explains how outputs from the project will be disseminated and identifies possible avenues for collaboration between the SERISS infrastructures, national statistical institutes and other producers of official statistics to ensure that the benefits of the SERISS project are felt beyond the immediate academic stakeholders and extend to the wider data community.

2. The SERISS project

The SERISS project, funded by the European Commission as part of its Horizon 2020 programme, is a four year collaboration (2015-2019) between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey. The project is coordinated by Rory Fitzgerald, director of ESS ERC.

The project (www.seriss.eu) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a larger role in informing European policymaking in future. It is doing this in three key ways:

- Addressing challenges for cross-national data collection, seeking ways to better represent the population and leveraging recent advances in translation;
- Breaking down barriers between social science infrastructures via training and networking events and the development of shared online tools to facilitate harmonised data collection and documentation;
- Embracing the future of the social sciences by examining the legal and ethical challenges associated with new forms of data, developing a cross-national probability-based web survey and exploring automated coding for socio-economic variables.

The project is committed to fostering links between the six SERISS infrastructures and other collectors, curators and users of statistical data including those involved with official statistics and surveys covering areas of the world beyond Europe itself. CESSDA already has links to the European Statistical System through the Data Without Boundaries (DwB) project and the SERISS Board of Strategic Advice includes representatives from organisations with close links to the ESS including CODATA and Eurofound. No doubt further opportunities for collaboration exist and will be developed as then project progresses.

3. Addressing challenges for cross-national data collection: Striving for quality

One of the five key areas of ESS Vision 2020 is “strive for quality”, intended to ensure that the ESS is guided by quality in all activities and continues to deliver “coherent, relevant and reliable statistics based on … sound methodologies …” (ESS, 2014, p7). This objective is also a key part of the SERISS project with a specific focus on the quality of cross-national social surveys.

The two leading social surveys in the SERISS consortium, the European Social Survey (ESS) and the Survey of Health Ageing and Retirement in Europe (SHARE), both follow an input harmonised model for design. Where possible, essential survey conditions such as sampling, mode of data collection, questionnaire format and data structure are kept the same across all participating countries in order to facilitate comparability. However it is not always possible or advisable to keep everything the same as this may force the adoption of the lowest common denominator across countries (Lynn, 2001). With sampling for example it is not always possible to use a frame of individuals for the selection of target respondents in all countries and alternatives have to be used. And with translation it has been found that in some countries and languages it may be advisable to move away from a translation that is grammatically close to the source questionnaire to achieve functional equivalence.

The SERISS project is focused on ways to improve data quality and increase confidence in the resulting statistics by addressing sampling and translation issues.

3.1 Sampling

All social surveys need to represent the population. Drawing a proper probability sample is therefore a central task both for academically led surveys and in official statistics. However, this is often difficult due to restricted access to sampling frames; the varying nature and quality of sampling frames; the trade-off between cost-effective deviations from simple random sampling and the size of the consequent design effects; and a climate of lower response rates. Furthermore, panel surveys require rather complex sample refreshment due to attrition. Work in this area is developing ex-ante and ex-post solutions. One strand of work involves mapping sampling practice across European surveys with the aim of exploiting synergies and economies of scale and possibly pooling sampling resources in future. A second involves assessing the potential for exploiting pre-existing administrative data to better understand and overcome survey non-response. The third strand looks at weighting for complex survey design and the fourth looks at handling item non-response including developing generally applicable imputation techniques. The final strand examines the feasibility of including the institutional population in general population sample surveys.

3.2 Translation

The production of high quality, comparable data in cross-national surveys depends on the production of equivalent language versions in each country. Despite significant advances and research into the efficiency of different approaches to translation assessment such as committee review meetings (Harkness and Behr, 2008) and back translation (Harkness et al., 2009), little attention has been paid to central aspects of survey translation such as how adaptation affects translation quality or how to best apply technological advances in translation studies to improve quality and reduce costs. There are three key areas of work. The first and largest involves the comparative testing of different questionnaire translation approaches. It aims to provide empirical evidence on best practices for country-specific adaptations in survey translation. The second looks at the feasibility of applying computational linguistic methods to survey translation and the third area involves the updating of the Translation Management Tool already used by the SHARE survey for use by other surveys.

4. Breaking down barriers: Promote efficiency in data production

Scarce resources and the pressure to do more with less are the reality under which producers of official statistics and academic surveys are operating. To address this challenge, a second key area for ESS Vision 2020 is to "promote efficiency in production processes" through "systematic collaboration" and "sharing knowledge…tools, data and services where appropriate". (ESS, 2014, p15). One way in which SERISS seeks to promote efficiency in production is through developing shared online platforms and software tools to support the data collection process.

Even the most sophisticated surveys still tend to rely on basic tools and processes with a lot of work being carried out manually, resulting in wasted or duplicated effort as well as scope for important information to be lost and inconsistencies to occur. The cutting edge tools developed under the SERISS project - and made available to other users - aim to streamline and harmonise the implementation of the survey lifecycle and to facilitate more efficient, standardised and transparent data collection across cross-national surveys.

4.1 Tools to support the survey lifecycle

A number of tools are being developed to support the documentation and realization of different stages of the survey lifecycle. These include an online Question Design and Documentation Tool (QDDT) and Question Variable Data Base (QVDB) which, together with the Translation Management Tool will generate and store metadata about different stages of the survey lifecycle. Other tools being developed include a web based survey management portal to provide a virtual collaborative workspace for multiple stakeholders on international projects and a data harmonisation platform which will facilitate more efficient and well documented output harmonisation by providing a platform via which data users can deposit and receive feedback on their harmonisation routines.

Many of the tools are being developed to be DDI-compatible to facilitate efficient generation and sharing of survey metadata across different stages of the survey lifecycle and different survey infrastructures. The tools and available metadata will be accessible not only to the primary data producers but also secondary users interested in finding out information about the data produced for subsequent analysis or replication on other surveys.

4.2 Automated socio-economic coding

Occupation, industry, employment status and educational attainment are core variables in many socio- economic and health surveys. They are also key parameters for official statistics, either on their own or to provide breakdowns of other statistics by sub-group. However, their measurement, especially in a cross-cultural, cross-national and longitudinal context, is cumbersome, not sufficiently standardised and often expensive. SERISS is therefore developing a cross-country harmonised, fast, high-quality and cost-effective coding module for these variables. This will be underpinned by large multi-lingual databases with tens of thousands of entries about job titles, industry names, fields of education and training, educational qualifications and employment status. Using search tree navigation or semantic matching techniques, interviewers will be able to code all this information directly with the respondent during the interview process rather than having to rely on recording sufficient verbatim information to enable post-coding. Once developed, the module and associated databases could be used on any survey including, for example, the Labour Force Survey to improve the speed and accuracy with which such important socio-economic variables can be coded.

The tools developed under SERISS will be made available to external users beyond the end of the project and will be accompanied by comprehensive training materials and thorough technical documentation to facilitate take up.

5. Embracing the future of social sciences: Harness new data sources ***

Official statistics are increasingly based not only on traditional surveys but also newer sources of data including administrative data, geospatial and big data as well as survey data collected via new methodologies e.g. online. A third key area for ESS Vision 2020 aims to harness these new forms of data. New forms of data and data collection methodologies open up opportunities but also raise new challenges including issues around access, quality and data security (Couper, 2013). SERISS, like the European Statistical System (ESS), is committed to equipping the social sciences for the future by exploring these issues and seeking to harness new forms of data alongside traditional survey data in a scientifically rigorous way.

5.1 Legal and ethical challenges associated with new forms of data

One important strand of work centres around exploring the legal and ethical challenges associated with harnessing new forms of data. Mapping exercises are currently underway to explore ways in which two increasingly important sources of data - social media data and administrative data - are being used by social scientists and to identify potential barriers to their access and use. Findings from the mapping exercises will be considered alongside the new European Data Protection Regulation to develop best practice guidelines to promote and facilitate the reuse of these new forms of data by social scientists.

Drawing on CESSDA’s expertise, SERISS also aims to develop new protocols for the secure storage and sharing of new forms of data to ensure optimal re-use. New workflows will be developed which account for the unique challenges for data curation posed by different forms of "Big Data" such as transactional and social media data, including their sheer scale and often transitory nature. A thorough understanding of the legal, ethical and practical considerations associated with the use and reuse of new forms of data is essential if these data are to be used in a robust and transparent way. Best practice guidelines will therefore be of particular interest to organisations looking to harness these new forms of data in official statistics.

5.2 A survey future online

Survey data collection is increasingly moving online. A number of national online probability based panels already exist including the LISS panel in the Netherlands, ELIPS in France and the German Internet Panel (Callegaro et al, 2014). One option for data producers and users interested in generating European comparisons is to attempt post-hoc output harmonisation of the data from these different online panels. However these panels have many different elements in their design and operation which have the potential to compromise comparability across countries. A second possibility is to set up a single probability-based web panel which operates cross-nationally and collects high quality input-harmonised data across countries. This approach mirrors the approach folloeed in studies like the ESS and SHARE with an input harmonized focus. Under SERISS, the European Social Survey aims to test a proof of concept for such a web-panel. A small probability-based panel will be established in three countries, the UK, Slovenia and Estonia, by drawing respondents from among those participating in the main face-to-face wave of the European Social Survey in 2016. They wil then be followed up monthly for one year.

The web-panel, CRONOS, provides an opportunity to explore different aspects of panel design and administration, including strategies for respondent recruitment and retention and ethical issues around offering incentives, providing devices to the offline community and secure storage of respondent records in a cross-national context. Measurement issues associated with online data collection will be explored including the prevalence of and strategies to overcome "bad" respondent behaviour such as straightlining and the extent to which differential attrition might undermine representativeness. Data will be made publically available. If successful, CRONOS may point the way toward a new mode of generating high quality cross-national data across a range of topics.

6. Opportunities for collaboration

This paper has identified areas of common ground between the academic social surveys and data infrastructures involved in SERISS and producers of official statistics. They face similar challenges if they are to "stay relevant", contribute to an informed society and feed into evidence-based policy making. The success of SERISS ultimately rests on the lessons learned and outputs generated by the project being successfully disseminated and exploited by other data producers and users, including those who form part of the European Statistical System and who are directly involved in producing data that is used by policy makers. Many of the issues explored under SERISS are important for the realisation of ESS Vision 2020.

There are a number of ways in which ESS members can engage with and shape the direction of SERISS and ensure that outputs produced will be fit for purpose. Stakeholder workshops - including one on using auxiliary data to address issues around survey nonresponse and another on the challenges associated with new forms of data - are planned and would benefit from the participation of ESS members. We are also seeking to establish a survey network bringing together data producers and users with different needs and expertise, including national statistical institutes, to learn from one another and from SERISS. Finally, SERISS is developing a comprehensive package of online and face to face training courses covering many different aspects of data management and analysis which may interest ESS members.

By forging links and sharing knowledge between established social science infrastructures, other social surveys and producers of official statistics, we can ensure the best possible social scientific evidence base with which to help address Europe's grand societal challenges.

7. References

Callegaro, M., Baker, R.P., Bethlehem, J., Göritz, A.S., Krosnick, J.A. and Lavrakas, P.J. eds., 2014. *Online panel research: a data quality perspective*. John Wiley & Sons.

Couper, M. (2013) Is the sky falling? New technology, changing media and the future of surveys. *Survey Research Methods* (2013). Vol. 7, No. 3, pp. 145-156.

European Statistical Service (2014) ESS Vision 2020 http://ec.europa.eu/eurostat/documents/42577/6906243/ESS+vision+2020+brochure/4baffcaa-9469-4372-b1ea-40784ca1db62 accessed 29th April 2016.

Harkness, J. A., Villar, A., Kephart, K., Schoua-Glusberg, A., & Behr, D. (2009). *Survey Translation Evaluation: Back Translation versus Expert Review*. Presented at the American Association for Public Opinion Research, Hollywood, FL.

Harkness, J. A., & Behr, D. (2008). *How instrument quality affects translation: Insights from a film of a team translation review*. Presented at the International Sociological Association Research Committee 33, Naples, Italy.

Lynn, P. (2001) Developing Quality Standards for Cross-National Survey Research: Five Approaches. *ISER Working Paper* 2001 21. https://www.iser.essex.ac.uk/files/iser_working_papers/2001-21.pdf accessed 29th April 2016.