# Project Outline: Data Integration: towards producing statistics by integrating different data sources

## Introduction

There are many new opportunities created by data sources such as Big Data and Administrative data. These sources have the potential to provide more timely, more disaggregated statistics at higher frequencies than traditional survey and census data.

National Statistical Offices (NSOs) face declining response rates and declining accuracy of survey data while at the same time there are increasing data needs at both the national and international levels. Despite tightening budgets, there are strong calls to NSOs to produce more timely data that is additionally ever more disaggregated. This will accelerate further with the needs for monitoring the Global Goals for Sustainable Development, Climate Change Conference and other multilateral agreements.

It is clear that NSOs are challenged by the capacities needed to incorporate new data sources in their statistical production process while at the same time companies have appeared exploiting these new sources to provide alternative statistics. If official statistics can't find an answer to this, we are at risk of losing our unique position. We can, however, join forces and keep or even increase our value proposition by providing relevant, reliable and comparable data of high quality. NSOs are particularly well placed to integrate data from various sources and to use them to satisfy the needs of policy makers and other partners for data. It is thus time to intensify our efforts and commence working on it within the framework of an HLG project.

## Project Objectives

The main objectives of the project are:

* to gain experience in data integration by pooling resources in joint practical activities.
* to translate experiences into general recommendations for data integration and to provide initial guidance for a quality framework

At least some of the activities will have to produce concrete outputs that can be replicated across different countries and/or be applied to produce statistics from integrating similar types of data sources. Creating an environment where organizations can jointly gain practical hands-on experiences would be an important outcome. However, project objectives go beyond merely gaining and sharing experiences. By creating vertical interlinkages in the process of identifying needs, selecting data sources, and methodological and quality considerations of the specific projects, the outcome as a whole should be greater than the sum of its parts.

## Project Scope

There are many possible types of data integration and for each, many possible combinations of data sources and modelling approaches. Only a limited number can be covered by a one year HLG project. The project will therefore be limited to a small number of concrete, practical examples that are fit for working in a collaborative setting. Consultation at the HLG Workshop and advice provided by the HLG executive committee will guide the prioritisation of the topics to include. The final selection will depend largely on the priorities of the participating organisations and the resources made available.

Many issues such as building partnerships, obtaining data; modelling approaches, quality, risks, comparability and metadata requirements will be explored in the individual integration activities. These experiences and experiences from past projects will be used to develop general recommendations for data integration.

## Content General Section

Activities will be grouped under five main types of integration: administrative sources with survey and other traditional data; new data sources with traditional data sources; geospatial data with statistical information; micro level data with data at the macro level; and validating data from official sources with data from other sources. The scope delimits the total number of integration activities under the project. Rather than adding packages, the five proposed packages will function as placeholders for all specific data integration activities that will take place under the project. The final size of the practical work packages (WP.1 to WP.5) may vary and some may be deferred or discarded depending on feasibility and interest.

Experiences from the HLG Sandbox have showed the benefits of working together on common data sets. But we have also learned that obtaining data for testing purposes can be a bottleneck. This will be dealt with by a specific Work Package (WP.0).

An integrating work package (WP.A) will be established to coordinate the cross cutting issues related to methodology and quality and to serve as a platform for sharing experiences. A summary of the experiences (including from earlier projects) could become the basis for a generalized methodology and quality framework.

As guiding tools, standards developed under umbrella of the HLG, such as GSBPM, GSIM and GAMSO will be used and where possible, CSPA compliant services will be developed or proposed.

### Work Package 0: Data sets for common approaches

Practical work in a collaborative spirit starts with data. If there are no suitable data sets that can be shared among people and institutes the best we can hope for is sharing of methods, tools and experiences. As soon as we are able to actually share data we can move to a next level of much more intense forms of collaboration. That is ultimately more rewarding. It is therefore important, and of value in itself, how and where we can make available data sets for common usage.

The "where" question is relatively easy to answer. It is now very likely that the Sandbox environment, set up at ICHEC in Ireland, will remain available in the near future. Although it was established for big data experiments, it is increasingly clear that the Sandbox may also be useful for other applications that require a versatile high-performance computing environment. This applies in particular to the Data Integration project, where computationally intensive methods will be required. One particular challenge is that the Sandbox is a relatively open environment that is not designed for dealing with privacy-sensitive or otherwise confidential data. For data integration purposes, in many cases the data to be studied will require some form of security.

The "how" question is a lot harder. How can we deal with the challenge to store and process data of a confidential nature, in such a way that it can be investigated in concrete multi-party tests and experiments? This may seem paradoxical, but in fact the Big Data project had to deal with similar challenges- and was able to solve them to achieve inspiring results. We can build on their experience, while we no doubt will have to be even more ambitious and we will have to push further.

The idea is to obtain data sets, make them available and prepare them so they can be used for testing data integration. Some kinds of data can probably be used without too much difficulty, like web-scraped data or social media sources, government-owned open data or public statistical outputs. In other cases, already available public use files or IPUMS data may be made available. A step further would be to systematically anonymize real data from surveys, censuses, government registrations or privately held big data sources. A final step would be to construct synthetic data sets that satisfy relevant characteristics of the original source they are supposed to mimic.

In all of the above cases, we have to consider various statistical, methodological, legal and ethical issues at the same time. This can become quite complicated but the fact that we only would need data sets for experimenting with data integration eases the burden. When suitable methods, processes and tools have been developed in the Sandbox environment in a collaborative way they can be moved to the secure environments of individual institutes for further testing on real data. Then the joint Sandbox experience will also prove helpful to support each other when additional issues appear. Data sets that are already available from the Sandbox will be the first to be considered for use in data integration.

#### Activities

* Identify and obtain data sets that can be used for testing in a collaborative environment
* Prepare and make data sets available in the Sandbox environment

### Work Package 1: Integrating Survey and Administrative Sources

This work package focuses on integrating administrative data with survey data.

Some countries have already worked on integrating data sources and there have been collaborative projects in this area, for example under Eurostat projects. These experiences will be analysed and used as a starting point for this project.

The administrative data may have existed for some time but not been used. It may be integrated using record linking or statistical matching or may use modelling approaches. It may involve pooling or combining information from multiple surveys, including surveys not conducted by the NSOs themselves.

The aim of this work package is to show methods of dealing with typical challenges faced in the integration of data from those sources. The quality of administrative dataset may be good enough for administrative purposes but not sufficient for statistical purposes. Transforming administrative datasets into statistical datasets may require improving the quality and dealing with conceptual differences, especially when we want to use administrative data in a direct way. In the case of surveys carried out with the use of data from administrative sources it is crucial to gather all data (from survey and administrative sources) in one database.

Examples of sources that can potentially be integrated are: Labour Force Surveys and social insurance register and/or educational registers, data from ministries of culture and cultural associations to produce statistics on museum attendance and there are several examples of administrative data being combined with survey data for producing indicators traditionally collected through censuses.

#### Activities

* Definition of statistical framework and boundaries of the work.
  + Administrative sources and a survey dataset have to be selected.
  + A case have to be defined: using records from the administrative dataset instead of records from a survey (a smaller frame) or adding additional variables from the administrative survey (a shorter a survey questionnaire) or both.
* Design expected outputs of the integration
* Design and test of transformation of the administrative dataset into the statistical dataset.
  + A cleaning method has to be designed
* Design and test of the integration of administrative data and survey data.
* Evaluation of statistical outputs obtained from the integration of administrative datasets and survey datasets.

### Work Package 2: New data sources (such as big data) and traditional sources

This package focuses on integrating data sources that traditionally have not been used by most organizations. One clear example of this is big data from mobile companies, social networks, smart sensors, satellite imagery, web pages, credit card transactions, etc.

Scanner data can be used to update Consumer Price Indices and information from Twitter and mobile phones combined with tourist and mobility surveys to create new or more frequent updates of statistics. A concrete example of an activity that is envisaged under this Work Package is to investigate how to integrate Internet-scraped data with traditional statistical datasets, namely administrative sources, surveys and statistical registers.

The relevant issues that will be investigated include:

* Definition of the statistical framework for using the Internet-scraped data, e.g. population, unit characteristics, quality metadata, etc.
* Entity extraction and recognition from the Internet-scraped data, according to the defined statistical framework.
* Object matching between Internet-scraped data and traditional statistical datasets.
* Design of statistical analyses starting from the result of the Object matching activity.

All the above issues will be investigated both theoretically and experimentally on a specific domain, namely Enterprises Web sites. This domain has the following interesting features:

* Limited privacy constraints.
* Variety of site structure, which allows experiments with different kind of techniques for processing Internet-scraped data.
* Amount of Web sites, which is very huge.

#### Activities:

* Definition of statistical framework and boundaries of the work. In this activity, the specific web sites and the specific information to extract from them will be identified. In addition, the traditional statistical datasets to integrate with such information will be selected. This process will be guided by specific statistical outputs for which the integration is relevant.
* Design and test of the integration between Internet-scraped data and traditional statistical datasets. In this activity includes both Entity extraction and recognition as well as Object matching activities will be carried out. New techniques will be explored, starting from the record linkage setting by extending it to the object matching one (i.e. where an object can have a looser structure than a record).
* Design and evaluation of statistical outputs obtained from the integration/combination/fusion of Internet-scraped data and traditional statistical datasets.

### Work Package 3: Integrating geospatial and statistical information

This package addresses the integration of geographic information with statistical information. Statistical data provides elements that national policy makers and international organizations need to make decisions and to develop public policies.

Statistical data is almost always related to a certain physical space, like a municipality, a State, a Country a region, etc., each level is useful for different actors and different kinds of decisions. Many of those decisions are conditioned by physical elements from the environment, and beyond that, they will have an impact on it. Location and amounts of natural resources, soil types, weather conditions, communications infrastructure, facilities, etc. are examples of geographic information which are indispensable elements to fully understand the figures that official statistical generates. Works made in the field of geospatial classification could be a departure point to link the fields of statistics and geography.

Integration of geographical data into the official statistical field aims to improve the value of the statistical information that is being produced.

#### Activities:

* A review of other efforts in the international context (like Eurostat’s GEOSTAT and GISCO and UN’s GGIM)  which are being carried on, to generate synergies and improve the capacities to work in the project
* An inventory of sources and projects related with geographical information which could be relevant to the production of statistics
* Development of proposals of valuable information which could be produced by integrating statistical and geographical information
* Research on methodologies to integrate and produce relevant information products derived from the combination of both kinds of data
* Conduct some experiments and pilots to assess the value of the integration of statistical and geographical data and to adjust the scope of the project

Due the complexity implied in this working package it is not expected to cover all the activities in just one year. It will just be the onset as chances are high that future activities in this area will demand efforts that will go beyond a single task team.

### Work Package 4: Micro-Macro integration

A specific area where data integration is useful concerns the reconciliation of statistics from accounting frameworks like National or Environmental accounts on the one hand, and figures from related primary statistics that have fed into these frameworks on the other hand. Due to the process of balancing, conceptual adjustments and other treatments it is often difficult, sometimes even impossible, to relate macro-figures and micro-figures that should in principle describe the same phenomena. There is a clear added value in joint analysis of such micro-macro statistics and interest in this area is growing. An area where already some progress has been achieved is that of the distribution of income and wealth for the household sector, including consumption patterns. When additional areas are being proposed in the international statistical community, the Data Integration project could support both the concrete practical work involved, e.g. by experiments in the sandbox environment, and the more abstract theoretical work, by gathering best practices for methods and processes and by suggesting ways to assess the quality achieved.

This type of integration might be of potential use for the many new challenges that come from the needs to monitor the Sustainable Development Goals that were recently adopted by all governments at the United Nations General Assembly.

It would build on the experiences of the most recent micro-macro project of OECD that was initiated by Eurostat and OECD; where information from SNAs and micro-level wealth and on income distribution statistics from household surveys as well as from administrative and private data sources are combined. If countries start activities under this package, OECD and Eurostat are willing to participate and to provide guidance.

#### Activities:

* Identifying potential sources and statistics that can benefit from micro-macro integration
* Road test existing guidelines in particular national settings and review, refine and further develop them

### Work Package 5: Validating Official Statistics

Another specific area where data integration comes into play is using data from other sources as part of validating official statistics. There have been cases where other sources are seen as comparable to official statistics, and when they differ, the official statistics have been challenged. One example from the United Kingdom shows how the distribution of businesses listed in the "Yellow Pages" telephone directories was compared to the coverage of the statistical business register (<http://www.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf>). A further example concerns comparisons of inflation figures from MIT's Billion Prices Project against official price indices. These examples show that "other sources" are reaching a level of credibility that challenges the role of official statistical organisations. Market researchers are struggling with similar issues. Their business is already under increasing pressure from cheap or free internet panels. One thing is clear. These "other sources" are here to stay and they will increase in number and influence. Several SDG will use indicators produced from official sources which may not use standardised methodologies and data sources. The different approaches may lead to discrepancies in the results which will have to be analysed and explained.

This work package will consider the issues involved in integrating alternate data sources into the validation processes used for producing official statistics. Issues include assessing origin and quality of the source, including trustworthiness and commercial or other interests of the parties exploiting them; designing processes and modelling techniques which are sustainable and formalised (as ad hoc adjustments to the statistics would be difficult to defend); and, educating users on proper use and interpretation of information (both the general public and more specific user groups).

#### Activities:

* Identify issues related to systematically using data from other sources in the validation of official statistics.
* Recommend potential approaches and modelling techniques.

### Work Package A: Synthesis Lessons Learned from new working methods

The problems with survey data and the new opportunities indicated in the introduction, led Citro to state that “We must move from a paradigm of producing the best estimates possible from a survey to that of producing the best possible estimates to meet user needs from multiple data sources”1. Our challenge is therefore to integrate diverse sets of inconsistent data. We will have to produce stable output with unstable ever changing inputs. The big change is that instead of trying to produce the best possible statistics from a single survey the reality will be a situation where we look at the information needs and try to find the best possible combination of sources to will deliver the best possible indicator/statistics that satisfied the users’ needs. Survey data might be part of the solution, but administrative data and Big Data and other non-traditional sources, will equally have to be considered. Additionally, instead of relying on one source, we have to consider all possible sources that can improve the usefulness and accuracy of our statistics. Integration can be at the micro level, at the level of a common denominator, at the aggregate (macro) levels, through modelling approaches or a mixture of these. Although there have been some attempts to integrate various data sources to produce statistics, no generalized methodology or quality framework exists. Given that it is an urgent and complex but common challenge, it is time to take this up jointly and to pool our resources to make a start with a more systematic approach towards developing a new common framework for statistical production.

With surveys we have the sampling error, a-selectiveness, representativeness and non-response. If we combine different sources, we still need an indication of what the data and statistics really measure, what the “uncertainty” or the reliability of the estimates are and how comparable the indicators are over time and with similar indicators produced from traditional sources. For integrated data we also need an indication of the quality and comparability if the sources change (either in quality or type).

To interpret the data and to assess the quality and 'uncertainty' of the estimates, adequate metadata has to be included. Quality frameworks have been developed for the whole process from operationalization to sampling, conducting the survey and processing the microdata and adjustments and corrections etc. Under the 2014 HLG Big Data project, a draft quality framework for Big Data was produced. These and traditional quality frameworks can be tested and used as a basis.

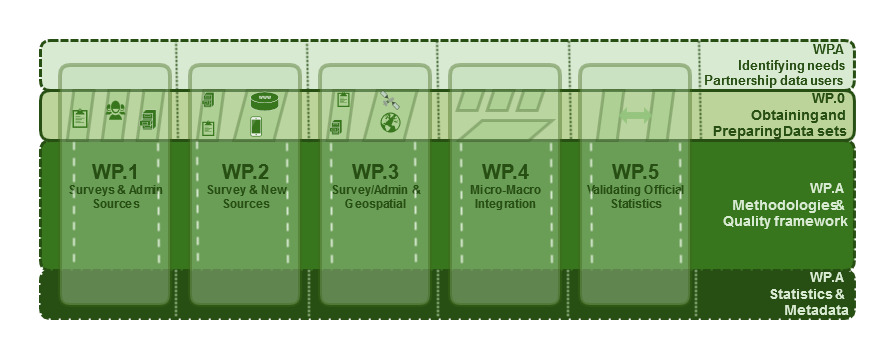
As guiding tools, standards developed under umbrella of the HLG, such as GSBPM, GSIM and GAMSO will be used. This might also lead to suggestions for adaptation of these standards to cover the new way of working under the new paradigm for producing statistics. For any services developed to integrate the data, CSPA compliance will be encouraged.

Methodology depends largely on the data sources used and there will not be a common way for integrating each type of data. This work packages will not attempt to develop this.

#### Activities:

* Recommend how to identify information needs and build partnerships with data users and producers
* Recommend how to assess and select data sources.
* Risk assessment of data sources
* Quality attributes of data sources
* Comparability issues in time-series and with statistics from traditional sources
* Recommend which metadata should be provided to each user group
* Develop a first draft of a quality framework for statistics produced from integrated data sources
* Identify common standards and frameworks to be used and provide guidance for improving them

It will not be possible to develop a final framework during the timeframe of the project. Nevertheless, we will aim to do more than merely a stock taking exercise of lessons learned from the practical work packages and from experiences from other projects . We will try to produce a foundation for the new paradigm in statistical production. It is essential that the activities of this work package are integrated into the practical activities in the project. Each of the practical work packages will include these topics and will be given opportunities to bring experiences together and to synthesise them into more general recommendations. This work package will ensure that experiences from one activity will be used in the other activities where relevant. Combining the practical experiences with more theoretical quality and metadata considerations should create synergy to improve the practical projects.



1 Citro, Constance F. (2014), From multiple modes for surveys to multiple data sources for estimates. Survey Methodology, December 2014 137 Vol. 40, No. 2, pp. 137-161. Statistics Canada

### Definition of Success

The five practical work packages are placeholders for possible specific data integration activities. The overall minimum success level for will therefore be reached if countries have been able to share experiences and have jointly gained experiences in integrating various data sources in WP-1 to WP-5. A high level of success will be indicated if also one or more concrete statistics are produced under one or more of these practical work packages.

An additional success criteria relates to WP-A. It will be met if experiences from individual projects can indeed be translated into more general recommendations and the experiences used as a starting point for a general methodological and quality framework for integrated data that needs to be developed in the following years. WP-0 will have been successful if data were obtained and made available in the Sandbox environment in a timely manner for use by the practical Work Packages WP-1 to WP-5.

### Expected Costs and risk factors

There is a wide interest both from NSOs as well as international organisations in this topic. Participating organizations are expected to make activities that take place under the project part of their work program. The table below shows an estimate of the resources and potential costs involved. The total amount depends largely on the total number of specific practical integration projects that will be undertaken. Not all work packages might attract sufficient support for an activity and others might attract multiple integration sub-projects. Risk factors are therefore either not enough support to start the activities under a work package or the reverse, too many activities. The feasibility of producing concrete output will be used to select the types of data integration while at the same time, more experimental integration projects will be considered.

| **Work Package** | **Resources**  **(estimate)** | **Source** | **Other costs** |
| --- | --- | --- | --- |
| WP-0 | 4 person months | Volunteer NSOs plus UNECE Secretariat | costs of Sandbox participation |
| WP-1 | 4 person months per sub-project | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP-2 | 4 person months per sub-project | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP-3 | 4 person months per sub-project | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP-4 | 4 person months per sub-project | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP-5 | 4 person months per sub-project | Volunteer NSOs plus UNECE Secretariat | Possible travel costs Workshop/Sprint |
| WP-A | 2 person months (mostly integrated in WP-1 to 5) | Volunteer NSOs plus UNECE Secretariat |  |
| Project Management | 6 person months | A project manager working with the UNECE Secretariat. Input from Executive Board and HLG members (in their role as project sponsors) | Travel costs for project events |
| Total | 32 person months  depending on number of sub-projects which proceed |  | US$ |

### Time Lines

The project will aim to complete the activities described by the end of 2016.

All work packages will start simultaneously and run for the full period of the project. WP0 will have to finish making data sets available before the other WP can start analysing them (it will continue to provide technical services to the other WP during the project).

### Governance

The project sponsor is the HLG. This is the group that has ultimate responsibility for signing off the project deliverables. In practice, this responsibility will be delegated to the Executive Board. A project manager will have day-to-day responsibility for the running of the project, providing regular updates and signalling any issues to the Executive Board as necessary.

Work will be mainly via electronic communication (email, wiki and webex). Face-to-face or virtual sprints will be considered when needed. Other meetings will be used to facilitate additional face-to-face meetings or to generate discussion. Guidance will be sought from the Modernisation Committees in the areas of their responsibilities.