

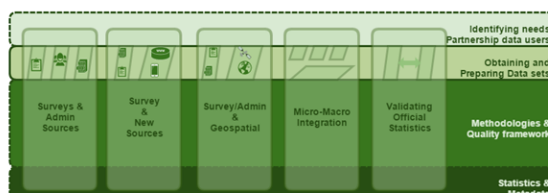
# WPA: a framework for Data Integration

[2016 ModernStats Data Integration Project Introduction and Purpose HL](#)  
[G-MOS Data Integration Project WPA: a framework for Data Integration WP0 : Data sets for common approaches](#)  
[WP1: Integrating survey and administrative sources](#)  
[WP2: New data sources and traditional sources](#)  
[WP3: Integrating geospatial and statistical information](#)  
[WP4: Micro-macro integration](#)  
[WP5: Validating official statistics](#)  
[Quality Framework for Data Integration](#)  
[Next Steps 2016](#)  
[Experiment Proposals and Reports 2016](#)  
[Project Members](#)  
[References and further information 2016](#)  
[Experiment Reports](#)

## 4. A Framework for Data Integration (WPA)

There are many possible types of data integration and for each, many possible combinations of data sources and modelling approaches.

Five common types of integration are: administrative sources with survey and other traditional data; new data sources (such as big data) 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.



### <sup>3</sup> Common Data Integration Types

The problems with using surveys as the most common approach to generating official statistics and the new opportunities arising from multiple data sources, led Connie 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”(Citro, 2014).

Our challenge is therefore to integrate diverse sets of inconsistent data and to produce stable outputs with often unstable, ever-changing inputs. Instead of trying to produce the best possible statistics from a single survey, we need to try to find the best combination of sources to deliver the indicator/statistics that best satisfy the users' needs.

Survey data, administrative data, big data and other non-traditional sources need to be considered. 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 a number of 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, the project aims to pool our resources to make a start with a more systematic approach towards developing a new common framework for statistical production.

## Broad Approach

There are many issues in undertaking data integration projects, such as identifying needs, building partnerships, obtaining data; finding appropriate modelling approaches, and managing quality, risks, comparability and metadata requirements.

The project has developed a set of broad topics to organise the issues to consider. These are:

- Business Requirements
- Opportunities
- Challenges
- Risk mitigation
- Standard Processes
- Recommended methods
- ICT considerations
- Quality
- Standards
- Metadata requirements
- Related work in other projects/organisations
- Skills
- Resources
- Partnerships
- Governance
- Promotion and advocacy
- Recommendations

Participants in the work packages are considering these topics as the practical experiments are undertaken, are generating lessons learnt and, where relevant, developing recommendations for this guide. As well, the project is bringing the experiences together, to synthesise more general recommendations which relate to more than one type of activity.

## Issues and Learnings

The following section summarises the general issues and learnings identified so far. Each work package has a similar list, although not all topics are necessarily relevant to all work packages. Topics are coloured “grey” where content has not yet been developed.

### Opportunities

There are a number of strong motivators for official statistics organisations to improve capacity for integrating data. Integrating different types of data can:

- provide more timely and more detailed statistics
- provide new official statistics
- meet new and unmet data needs
- lower response burden
- overcome the effects of reducing response rates
- address quality and bias issues in surveys.

### Challenges

There are also many challenges including:

- the different forms of data integration
- producing stable output with unstable inputs
- the need for new skills, new methods and new information technology approaches
- the need for new concepts or aligning existing statistical concepts to the concepts in new data sources
- the need to identify the minimum and ideal metadata required
- measuring, managing and publishing the quality of both the data sources and the statistics produced
- forming effective partnerships with data providers, commercial companies, academia and others
- moving data integration projects from research projects to repeatable, reliable production of statistics
- setup and ongoing costs
- the speed of adoption
- governance for data integration projects
- managing public perception and communication
- avoiding duplication of effort across countries and organisations and using the collective experience of the official statistics community.

### Risk mitigation

Factors to consider include:

- use of organisational risk management frameworks with data integration projects
- use of Guidelines on Risk Management Practices in Statistical Organizations (currently being prepared by Istat)
- specific risk assessment approaches for data sources.

### Standard Processes

Many data integration projects could follow a similar series of steps, such as:

- identifying needs and clarifying business requirements
- researching related work done in other organisations
- identifying partners and collaborators
- selecting potential data sources
- identifying methodological and quality considerations
- analysing whether data sources can be used (as direct sources or for validation)
- making a business case (including costs, benefits, risks, etc)
- obtaining the data
- obtaining required tools, skills, resources
- experimenting
- assessing results
- refining methods and approach (as required)
- developing into a repeatable production solution.

The project plans to explore the common steps required and relate these to the Generic Statistical Business Process Model (GSBPM) for the next version of the guide.

### Recommended methods

Methodology depends largely on the data sources used and there will not be a common way for integrating all types of data. However, the project is keen to form a catalog of commonly used and new methods for the different types of data integration. Ideally, the methods will be described in a form that is compatible with the Common Statistical Production Architecture ("CSPA - UNECE Statistics Wikis,")

## ICT considerations

The project has access to the Sandbox environment set up at Ireland's High Performance Computing Centre (ICHEC) for the purposes of running experiments. ICHEC was established for big data experiments but is also useful for other applications that require a versatile high-performance computing environment, such as developing common approaches for data integration, where computationally intensive methods may be required.

Often the tools available in the Sandbox environment are not currently available in official statistics organisations.

## Quality

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 to develop a first draft of a quality framework for statistics produced from integrated data sources.

## Standards

- As guiding tools, standards developed under the umbrella of the HLG, such as GSBPM, GSIM and GAMS0 will be used and where possible, CSPA compliant services will be developed or proposed.
- We also need to identify other common standards and frameworks used in the wider environment that are relevant to data integration tasks.

## Metadata requirements

Through the work on the individual experiments, the project is looking at minimum and preferred approaches to metadata.

## Related work in other projects/organisations

We have gathered information about related work in other projects and are keen to link to this work wherever relevant. Ultimately, we expect this guide to provide a starting point from which the user can find information about recommended approaches that already exist or are currently being developed.

## Skills

## Resources

## Partnerships

Partnerships often need to be built with data users, the general public, data producers, academia, software and system providers, other government, international and commercial organisations.

## Governance

A number of countries have developed strong governance frameworks for managing data integration projects. For example, Australia has a Statistical Data Integration Framework ("National Statistical Service § Data Integration Need to know," 2014) which has been endorsed and is used across the Australian government. Other countries have similar frameworks which will also be referenced in future versions of the guide.

## Promotion and advocacy

Two of many issues in this area are:

- Using this guide as it is developed to encourage official statistics organisations to use and contribute common approaches in pursuit of more efficient and effective adoption of data integration into their production processes.
- Promoting understanding and support of data users and the general public of the data integration work being done within official statistics organisations