**UNECE High-level Group for the
Modernisation of Official Statistics**

**Project Proposal: Develop a practical online Guide to Data Integration for Official Statistics and undertake specific experiments on priority areas**

|  |
| --- |
| This project proposal was prepared by the Data Integration Project and is submitted to the HLG-MOS for approval. |

|  |
| --- |
| **1 Purpose** |
| Data integration provides the potential to produce more timely, more disaggregated statistics at higher frequencies than traditional approaches alone. In 2015, HLG MOS recognised that official statistics organisations were challenged by the capacities needed to incorporate new data sources in their statistical production processes. The 2016 Data Integration project was commissioned to * Gain experience in data integration by pooling resources in joint practical activities.
* Translate experiences into general recommendations for data integration and provide initial guidance for a quality framework.

It was recognised that it would not be possible to cover all types of data integration in a single year. After a slow beginning, the 2016 project gained momentum with increasing involvement from member countries over the year. The project:* Identified and jointly progressed a number of practical experiments of priority interest to participating organisations.
* Learnt lessons from these experiments and other related work to develop initial recommendations.
* Designed the initial structure and content for a guide to data integration for official statistics.

For 2016, the project was structured into 7 work packages: WP0: Data sets for common approaches WP1: Integrating survey and administrative sources WP2: New data sources (such as big data) and traditional sources WP3: Integrating geospatial and statistical information WP4: Micro-macro integration (inactive in 2016) WP5: Validating official statistics WPA: Synthesize lessons learnt from new working methods.11 experiments were designed, based on priority work occurring in participating countries. The project has found significant common interest in progressing the experiments further and in developing a practical guide for future projects. For 2017, the project proposes to:* **Develop an online, adaptive, practical guide to Data Integration for Official Statistics** which supports successful data integration projects; using lessons learnt within the project and in related work.
* **Undertake further joint experiments in high priority practical interest areas**. The project has identified a number of areas where working together should bring faster results than working alone.
 |
| **2 Project description** |
| **WPA: Develop an online, adaptive, practical guide to Data Integration for Official Statistics (WPA)**Using the lessons learnt from the project experiments to date and from other initiatives, there is sufficient material to develop a practical guide for undertaking data integration projects. The main challenge is to organise this knowledge into simple, easy to consume and adaptable guidance to support successful projects and common approaches. Initial guidance for a framework has been created with these aspects identified as important for achieving success in data integration projects:

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

The guide should:* Cover how to get started on a data integration project and describe suggested processes to follow, based on GSBPM.
* Provide advice for obtaining datasets, forming partnerships, assessing quality and securing ongoing access (WP0).
* Outline the different types of data integration (as identified in the 2016 work packages) including specific issues and links to related work, standards and case studies.
* Provide lessons learnt and recommendations; with emphasis on the methods, the best ways to assess quality, and harmonizing and comparing outputs.
* Promote use of ModernStats standards and approaches (eg CSPA to describe methods/ICT solutions and GSIM to describe data).
* Provide a catalog of methods described using the Modernstats standards
* Distil experience within the project and in related work into additional guidance points.
* Be available online (probably using the UNECE Wiki system) and support easy adaptation as new experiences are added. Contributions (with attribution) should be sought from any related projects and from the various modernisation groups.

**WP0-WP5 Further work on joint experiments in priority interest areas**Participants in the project in 2016 are enthusiastic about further developing common approaches in particular areas of statistics to accelerate the use of multiple data sources and provide material for the guide. The following proposals have been developed within the project and by others interested in participating in 2017. Additional proposals (and participants) may arise at the HLG MOS Workshop. Priority should be given to those proposals which have clear deliverables, provide practical experience and provide practical material for the guide. **Proposed Activity: Align approaches for applying new data sources to integrated price measurement (WP0 and WP2)**Many countries have similar challenges improving the quality of the CPI in terms of coverage and real-time quantity. For example, ABS is using scanner data in production and planning to go to use of full-coverage soon; CBS Netherlands is actively researching use of online data in production and has been using scanner data in production; Statistics Canada is considering using Price Stats data in production, including use of an API; and Statistics NZ is using online/scanner data in production. Several data providers operate in many countries and there is an opportunity to develop a common approach that can be used in multiple countries. For 2017 the proposed activities are to:* Extend the range of products and data sources
* Organise an event to bring data providers and NSOs together to determine mutual benefits and develop agreements for data supply and production
* Test comparability of data across countries and time
* Develop ICT skills to receive load and explore data into the sandbox
* Define common methods required
* Coordinate with related work (e.g. by UN, IMF etc.)

**Proposed Activity: Create synthetic datasets for sandbox experiments (WP0)** The creation and documentation of a set of synthetic datasets for the sandbox will allow countries to collaborate on developing common methods, removing issues of confidentiality and encouraging use of the same data formats. This activity will incorporate the use of outputs from the Big Data Project already lodged in the Sandbox. It may also provide a practical base/examples of the proposed CSPA Data Architecture.**Proposed Activity: Develop practical guidance for integrating survey, administrative data and big data (including case studies) (WP1 and WP2)**During 2016, participants in these work packages investigated current work in this area (within their own organisations and in other groups (eg ESSNET and the HLG Big Data project), with specific interest focussed on job vacancies statistics, employment registers and labour force, and geographic location of schools. Lessons learnt, methods used, challenges and results were gathered and the group is preparing an overview of work and developing initial draft guidelines. For 2017 the proposed activities are:* Use the work being done within participating organisations to develop a practical guide to integrating survey, administrative data and big data (including case studies)
* Encourage the involvement of other participants and projects
* Describe the steps needed using the GSBPM.

**Proposed Activity: Develop practical guidance on integrating geospatial and statistical information (WP3)**The geospatial and statistical data integration landscape is very complex, with many players globally.  The Global Statistical Geospatial Framework (GSGF - UN GGIM), the Statistical Spatial Framework (SSF) and initiatives such as GEOSTAT2 (Eurostat) are vital for developing a consistent and systematic approach to linking geospatial and statistical data.  This is likely to take some time and a considered and organised approach is being pursued. However implementation of “The 10 level model” into the lower layers of GSGF will support practical activities of integrating statistics with geospatial data. The following activities are proposed:* Develop a decision tree - a path of practical dynamic questions to be answered before embarking on integrating geospatial data with statistical data. This will assist organisations to assess their maturity and capability for spatial statistics.
* Examine and consider incorporating the work done by participants of the GEOSTAT 2 project and Australia to reference geospatial capabilities and data standards (including standards from the geospatial community) in the components of the components of GSBPM and GSIM.
* Examine and consider joint work with UNGGIM, Europe and Americas, in order to obtain better results in integration of statistical and geospatial information taking as base advances obtain by this initiative.
* Compare and analyse the results of the surveys conducted in different regions. Conduct additional surveys based on GEOSTAT2 as required if methodologies of surveys mentioned above are different and results are not comparable. Analyse the existence of common points and divergence of approaches at the national and international level of spatial objects used in statistics and geodesy to harmonise this two reference frameworks (implementation “The 10 level model” into the lower layers of GSGF).
* Identify common risks for integrating geospatial and statistical data.

Because of the complexity of this area, it would be useful to conduct a face to face or virtual sprint of key players (including Australia, Poland, Colombia, Mexico, Finland, Sweden and others to be determined) possibly back to back, or incorporated into, a proposed workshop on geospatial and statistical standards (Sweden)**Proposed Activity: Develop practical guidance on using additional sources to validate official statistics (WP5)**Work in 2016 focused on identifying different applications and methods for validating official statistics. Issues identified with use of administrative data to validate official statistics and lessons learnt from experiments as well as other validation projects carried out within organisation of contributing members are documented to provide initial guidelines in the use of administrative data to validate official statistics as well as recommend approaches and modelling techniques to resolve issues identified.We propose, for 2017, to work towards:* Investigating the relevance of the ESSNET Rules Repository and ESSVIP validation definitions handbook as tools for validation
* Testing recommended approaches and modelling techniques
* Expanding the guidelines developed in 2016, if required. Include quality indicators that are essential to report on the quality of the validation process. Communicate findings with the HLG project developing quality indicators for the GSBPM.
 |
| **3 Alternatives considered** |
| 1. Many surveys and inventories have been created for HLG related initiatives in the past, providing useful information for anyone researching approaches. This proposal aims to take a step further to provide clear and simple guidance on what to consider when undertaking a data integration project, linked to relevant initiatives and underpinned by contributed case studies.
2. In particular areas, such as integrating geospatial and statistical information, there are international efforts to find and advocate common approaches. This proposal aims to support and link to these endeavours, not to replace or reinvent them.
3. If we do nothing, an opportunity will be missed to leverage the work done in the Data Integration project to date as well as many individual efforts within our organisations and by international groups. The current landscape for data integration is very complex and this proposal offers a way to provide a simple entry point and guidance for undertaking data integration projects. A small investment in 2017 has the potential to develop into an ongoing community of interest and guidance for data integration.
 |
| **4 Expected Benefits** |
|[x]  Reduced costs |
|[x]  Increased efficiency |
|[x]  Reduced risks |
|[x]  New capabilities to meet user needs |
| Justification: * A well-developed guide should improve the execution of data integration projects, by helping projects to consider all important aspects and to incorporate the experience of others.
* The guide will provide a structured focal point for sharing and reusing expertise and information
* The work will apply the Modernstats standards in practical activities.
* By agreeing on a structure for the guide and encouraging submissions, we can unlock expertise within our organisations to share with the official statistics community.
* Progressing joint work in priority interest areas will assist participants to more quickly develop those new capabilities and will ensure that the advice in the guide is based on real world experiences.
 |
| **5 Which key priorities in the HLG-MOS Strategic Framework does the proposed project relate to?** |
|[x]  Take cost out of our organisations to reinvest in more value added areas |
|[x]  Explore new areas collectively and leverage each other’s' research investments in specific areas |
|[x]  Provide whole of government data ecosystems based on international standards, for better estimates in key policy areas |
|[x]  Renew our governance and operating processes |
| Justification: * The 2016 Data Integration Project found opportunities to share lessons learnt and ways to act collectively on similar projects (eg. online and scanner data for integrated price measurement).
* Progressing joint work as proposed will leverage the research investment’s made by each other
* Directing data integration efforts towards using the Modernstats standards will support the development of whole of government (and in some cases international) data ecosystems, ultimately leading to better estimates in key policy areas.
* The groups in the new governance model (Blue Skies thinking, Supporting Standards, Processes and Skills, Sharing Tools) should be asked to contribute to the specific experiments; e.g. by providing advice, education and mentoring.
 |
| **6 How does the proposed project relate to other activities under the HLG-MOS?** |
| This proposal offers an opportunity to guide organisations undertaking data integration projects towards concrete, real world use of the ModernStats standards. As the project is based on practical activities, it provides an opportunity to demonstrate the use of the standards in real world activities and to feedback suggestions for future versions of the ModernStats standards. |
| **7 Proposed timetable** |
|  | Start: January 2017 End: December 2017 |
| **8 Expected resources and costs** |
| **WPA** **Develop and promote an online practical Guide to Data Integration for Official Statistics**Volunteers from NSOs to work together on common areas, to contribute case studies and content and to assess and critique the usefulness of the guide. Volunteers from modernisation groups (Blue Skies thinking, Supporting Standards, Processes and Skills, Sharing Tools) to contribute to the work packages and the guide in their areas of expertise6 person months for Project Manager/Editor for GuideCosts associated with 12-16 participants to attend Integrated Price Measurement and Integrating Geospatial and Statistical Data Sprints (if possible back to back with other meetings)Costs associated with 10-12 participants to attend Face to Face project sprint in 2017. |