

WP2: New data sources and traditional sources

A Guide to Data Integration for Official Statistics: High Level Group for the Modernisation of Official Statistics Data Integration Project

[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](#)

7. New data sources (such as big data) and traditional sources (WP2)

Coordinator: Tiziana Tuoto (Italy)

This package focuses on integrating new data sources (such as big data from mobile companies, social networks, smart sensors, satellite imagery, web pages, credit card transactions, etc) with traditional sources.

Activities:

- Define a statistical framework and boundaries of the work
- Design and test the integration between Internet-scraped data and traditional statistical datasets. Entity extraction and recognition as well as Object matching activities will be carried out. New techniques will be explored for record linkage and object matching (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.

Experiments:

- Web-scraping strategy case studies: Approaches to gaining web-scraped data for official statistics – case study of NZ and survey of other countries' experiences.
- Integrated big-data price measurement: Estimation and comparison of price indexes from different big-data sources across countries
- Integrating web scraped data for the compilation of price statistics
- Integrating potential information sources for the statistical data production on job vacancies

Broad Approach

During 2016, participants in this work package investigated current work in this area (within their own organisations and in other groups (eg ESSNET and the HLG Big Data project).

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

Issues and Learnings

The following section summarises the issues and learnings identified so far for this part of the project.

Opportunities

- Better sample, full coverage (e.g. the whole circle of products available online) <--> selectivity reduce respondent burden
- The opportunity to create new measures: cheaper and of higher quality
- Cheaper than survey data
- Advantages of online data: fast, high frequency,

Challenges

- We cannot be exhaustive
- A wide interpretation of the term data integration: we should be very explicit
- Wide range of problems we face using big data
- Wide range of quality

Risk mitigation

- Focus on what we expect from using big data, and on the outputs
- Validation through comparing big data with other data sources
- Interrupted data sources and the need for contingency plans
- Focusing on limited resources, on tangible and useful outputs
- Good understanding and documentation of data sources, to understand how to calibrate data

Standard Processes

Recommended methods

- Hedonic methods for price indexes
- FEWS index

ICT considerations

- Processes and systems need to evolve
- New IT skills
- The need for new computer applications and servers (proper IT background) for processing big data see Work Package 0
- Use of common area for the lodgement and storage of data (i.e. the sandbox for WP0);
- need secure and efficient file transfer mechanism to get data for production

Standards

- The lack of well-established standards: the opportunity to propose new standards

Related work in other projects/organisations (references)

- UNECE big data Eurostat big data
- The prices work (WP0)
- New Zealand rental index from trade me site (not quality adjusted)
- Job vacancies (Serbia)

Skills

- Knowledge of web-scraping tools
- Subject matter expertise (knowledge on price indices)
- Analytical skills
- IT skills
- Methodological skills

Resources

- Time
- The resulting datasets from WP0
- Budget
- Human resources with the necessary IT skills
- Specific IT resources (sandbox)

Partnerships

- Ensure we coordinate with international groups (UNECE and OTTAWA) on price indices and Big Data
- Developers/administrators of sandbox
- Data providers (WP0)
- Different agencies that will work on the data

Promotion and advocacy

- Among professionals interested in this issue (IT experts, methodologists, price index experts, business magazines)
- Commercial markets
- Back to data providers to show benefits of their data
- Promotion within in-house agencies
- Conference papers

Other factors

- Establish the outputs of the wider project
- To elaborate a timetable (if necessary or sufficiently flexible)
- Produce conference papers to start promoting the initial results
- Summarizes the technical difficulties that arose during the research
- Lodge code for methods in sandbox
- Identify gaps in classifications or needs for new classifications e.g. data integration

Recommendations

- Keep your objectives clear
- Start the work on a sample
- Look at the data before commencing
- Consult with other experts, find out more on possible methods and solutions in a practical sense
- Patience and persistence
- Have in mind a wide range of solutions

- Specify the characteristics of the problem to be resolved

Further Work

- Pointing out the mutual benefits – commenting them about possible modifications in their methods to get more accurate price indices
- Give feedback to data providers on the results
- Seek inputs on the methodology from other price index experts
- To expand the set of products involved in the experiment
- Thinking on this experiment, refine the methodology with existent data