**The Role of Big Data in the Modernisation of Statistical Production**

**I. Background**

1. The importance of the relationship of Big Data to the official statistics industry has been raised in a number of arenas during recent years. At a High-Level Seminar on Streamlining Statistical Production and Services, held in St Petersburg, 3-5 October 2012, participants called for a strategically-focused document aimed at heads and senior managers of statistical organizations, outlining the issues, challenges and opportunities that Big Data poses for official statistics. The resulting paper[[1]](#footnote-1) discussed definitions and sources, and identified challenges in the areas of legislation, privacy, financial aspects, management, methodology and technology. It suggested that there are a great many opportunities for the use of Big Data, broadly dividing these opportunities into three categories: combining Big data with official statistics; replacing official statistics by Big data; and filling new data gaps.

2. Subsequently the April 2013 meeting of the UNECE Expert Group on the Management of Statistical Information Systems (MSIS) once again identified Big Data as a key challenge for official statistics, and called for the High-Level Group for the Modernisation of Statistical Production and Services (HLG) to focus on the topic in its plans for future work. A temporary task team composed of representatives of 13 national and international statistics organizations was convened to formulate the present project proposal.

3. This project is important for the HLG’s broad programme of modernisation of statistical production. As a component of the modernisation programme, it will contribute to the goals of international harmonisation and collaborative approaches to new challenges, improved efficiency of statistical production, and the modification of products and production methods to meet changing user needs. The HLG’s strategy document[[2]](#footnote-2) states that “products and services must become easier to produce, less resource-intensive, and less burdensome on data suppliers” and that “new and existing products and services should make use of the vast amounts of data becoming available, to provide better measurements of new aspects of society”. The project is aligned with these aspirations since it focuses on new sources, new methods, new outputs, and ways to tackle the issues surrounding these.

4. This project outline includes the objectives, scope and content of this project, as well as some practical project management issues.

**II. Project objectives**

5. The project has three main objectives:

* To identify, examine and provide guidance for statistical organizations to identify the main possibilities offered by Big Data and to act upon the main strategic and methodological issues that Big Data poses for the official statistics industry
* To demonstrate the feasibility of efficient production of both novel products and ‘mainstream’ official statistics using Big Data sources, and the possibility to replicate these approaches across different national contexts
* To facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

**III. Scope**

6. This project concerns the role of Big Data in the modernisation of official statistical production. It will tackle strategic and practical issues that are multi-national in nature, rather than those that are specific to individual organizations or national sources. It will not attempt to identify a comprehensive list of all possible sources or uses of Big Data, nor can it hope to ascertain all the issues and challenges, let alone solve them, since these are broad-ranging and constantly evolving. While the project does involve a practical component and a consideration of methodological issues, its aim is not to focus on the technical details of analysis of Big Data, which are covered by other national and international projects, unless these are sufficiently cross-cutting to be of concern internationally.

7. By including representatives of many national and international statistical organizations in the task team that formulated this project proposal, and by continuing to consult with these and other partners throughout, the project aims to be complementary to other initiatives and to avoid duplication of efforts. It also aims to be as relevant as possible to organization-specific needs and concerns. The project itself will endeavour to include inputs from academia and the private sector in addition to the official statistics community, in order to maximise learning across fields and avoid ‘reinventing the wheel’.

**IV. Contents**

8. This project comprises the four work packages outlined below. As a precursor to the project, the following activities have been undertaken by the temporary task team and the UNECE secretariat:

* Formulation of a classification scheme for Big Data sources and identification of the attributes of these sources that are relevant to their use in the production of official statistics
* Development of a repository with examples of sources being used, products being created and other activities being undertaken by statistical organizations, according to the classification and attributes identified above. This repository can be viewed as a repository of case studies for organizations intending to use similar sources or undertake similar projects
* Initial specification of the ‘sandbox’ environment described under work package 2 below.

***Work Package 1: Issues and Methodology***

9. This work package involves an analysis of the major strategic questions posed by the emergence of Big Data. It will require, first of all, more concrete definitions of the various terms. The work package will require very broad inputs from across the statistical community and hence will begin with gathering input through electronic consultation and virtual meetings.

10. The work package will expand on, and seek to address the major challenges listed in the HLG paper ‘*What does Big Data mean for Official Statistics?’:*

* Legislative: how to access and use data?
* Privacy: how to manage public trust and acceptance of data re-use and linking to other sources?
* Financial: what are the potential costs and benefits of using Big Data?
* Management: what policies are necessary for the effective management and protection of the data?
* Methodological: how to manage data quality? Are current statistical methods and models suitable for Big Data?
* Technological: what are the issues related to information technology?

11. It will also address a variety of issues and questions identified by the task team, including (*but not limited to*) the following:

* How can we assess the suitability of Big Data sources for the production of official statistics?
* How can we effectively capitalise upon the promise of massively increased timeliness offered by many Big Data sources?
* Can we identify best practices or guidelines for the major methodological issues relating to Big Data? E.g.:
  + Methods for reducing data volume
  + Methods for noise reduction
  + Methods for ensuring confidentiality and avoiding inadvertent disclosure
  + Methods for obtaining information on statistical concepts (text mining, classification methods, etc.)
  + Methods for determination of population characteristics, e.g. determining the population of users of social media services through analysis of words or phrases that are highly correlated with certain demographic characteristics
  + Assessing the applicability of models
* Should Big Data be treated as homogeneous, or do they require different treatment according to the role they play in the production of official statistics?
  + Experimental uses
  + Complementing existing statistics e.g. benchmarking and validity checking;
  + Supplementing existing sources, permitting the creation of entirely new statistics;
  + Replacing existing sources and methods
* Are there ‘quick wins’, applicable beyond Big Data, such as data storage, technology, advanced analytics, methods and models which could transform our thinking in relation to the production of official statistics more generally?
* How should statistical organizations react to the novel idea that in a Big Data world there are no ‘bad’ data (they all tell us something)?
* How can organizations mitigate the risk of a data source ceasing to exist, or changing substantially, when it is outside the control of the organization?
* How can Big Data be combined with survey data? And relatedly, how can the transition from statistical data production based entirely on surveys to production based substantially on Big Data be managed?
* Do we need a research question before exploring a Big Data source, or should we just experiment and innovate to see what is possible?
* What becomes of the time series in a world where data sources and uses may become more transient?
* What is the demand for new types of statistical information, given the new possibilities?
* How should statistical organizations approach the need to ‘educate’ (or re-educate) staff and users?
* How will institutional structures need to change in order to support the use of Big Data and ensure its quality and the quality of resulting outputs?

12. The output from this work package will take the form of recommendations, good practices and guidelines, developed through broad consultation of experts throughout the official statistics community, and coordinated by expert task teams. The material will be collated in an electronic environment such as a wiki. Such an environment will allow the guidelines to function as a ‘living document’, permitting timely updating as circumstances change. The task of maintaining the content after its initial formulation will be overseen by the HLG’s Modernisation Committee on Products and Sources.

***Work Package 2: Shared computing environment (‘sandbox’) and practical application***

13. This work package will form the practical element of the project, aimed at proving concepts in two related strands:

(a) Statistics:

* the possibility of producing valid and reliable statistics from novel sources, including the ability to produce statistics which correspond in a predictable and systematic way with existing ‘mainstream’ products, such as price statistics
* the cross-country applicability of new analytical techniques and sources, such as the analysis of data from social networking websites. This will be done by attempting to reproduce the results of a national project in other countries

(b) Tools:

* the efficiency of various software tools for large-scale processing and analysis
* the applicability of the Common Statistical Production Architecture (CSPA – under development) to the production of statistics using Big Data sources.

14. A web-accessible environment for the storage and analysis of large-scale datasets will be created and used as a ‘sandbox’ for collaboration across participating institutions. One or more free or low-cost, internationally-relevant datasets will be obtained and installed in this environment, with the goal of exploring the tools and methods needed for statistical production and the feasibility of producing Big Data-derived statistics and replicating outputs across countries. Simple configurations with tools and data will, whenever possible, be released in ‘virtual machines’ that partners will be able to download in order to test them within their own technical environments. The details of the sandbox will be specified in a separate annex to this proposal, following a study of alternative scenarios and a consideration of criteria by a task team of experts.

***Work Package 3: Training and dissemination***

15. This work package will ensure that the conclusions reached in the two preceding work packages are shared broadly throughout the statistical world and beyond. This will be done through a variety of means, including:

(a) establishing and maintaining a central location and online infrastructure for documentation and information-sharing on the UNECE wikis, including detailed documentation arising from work packages 2 and 3

(b) preparation of electronic demonstrations of tools and results, for example in the form of Webex presentations and Youtube videos which can be disseminated widely. Identification of existing electronic resources and online training materials is also included in this strand

(c) a workshop in which the results of work package 2 will be presented to members of the various of expert groups involved in the HLG’s modernisation programme. This would be held back-to-back with the annual workshop on modernisation of statistics at which all these expert groups are represented (likely to be November 2014).

***Work Package 4: Project management and coordination***

16. This work package comprises the necessary project management activities to ensure the successful delivery of the other three work packages.

**V. Definition of success**

17. Overall, this project will be successful if it results in an improved understanding within the international statistical community of the opportunities and issues associated with using Big Data for the production of official statistics. Success criteria for the individual work packages are:

* Work package 1: a consistent international view of Big Data opportunities, challenges and solutions, documented and released through a public web site
* Work package 2: recommendations on appropriate tools, methods and environments for processing and analysing different types of Big Data, and a report on the feasibility of establishing a shared approach for using Big Data sources that are multi-national or for which similar sources are available in different countries.
* Work package 3: exchange of knowledge and ideas between interested organizations and a set of standard training materials
* Work package 4: the project is completed on schedule, and delivers results that are of value to the international statistical community.

**VI. Expected costs**

18. The following table shows an estimate of the minimum resources and other costs needed to deliver the different work packages. Each organization involved in the project will be expected to cover the costs of their participation (including wages and any travel expenses for participants).

|  |  |  |  |
| --- | --- | --- | --- |
| **Work Package** | **Estimated resources** | **Source of resources** | **Other costs (in US Dollars)** |
| 1: Issues and methodology | 8 person months | Volunteer NSOs plus UNECE Secretariat | Possible travel costs if a workshop or sprint session is needed |
| 2: Shared computing environment & practical applications | 12 person months | Volunteer NSOs plus UNECE Secretariat | Costs associated with renting a shared space and acquiring data and tools (max $10,000?)  Possible travel costs if a workshop or sprint session is needed |
| 3: Training & dissemination | 4 person months | Volunteer NSOs plus UNECE Secretariat | Up to $1,000 for costs associated with preparing and disseminating training materials and running workshop for expert groups |
| 4: Project management | 6 person months | A project manager working in the UNECE Secretariat. Input from Executive Board and HLG members (in their role as project sponsors) | Up to $500 for telecommunications and other incidentals  Travel costs for project events |
| **Total** | **30 person months** | **UNECE Secretariat (9 person months)**  **NSO / International organization staff (21 person months)** | **Up to $11,500 total costs as described above, plus:**   * **possible consultancy costs** * **travel costs of experts** |

**VII. Timetable**

19. The project will aim to complete the activities described by the end of 2014. There are, however, various unknowns which may affect the timetable:

* The availability of resources from national and international statistical organizations to support this project – if the necessary resources are not available, either the timetable will need to be extended, or the outputs will need to be re-defined (in terms of quality or quantity or both)
* The availability of project management and support resources in the UNECE Secretariat – to meet the resource requirements of this project will require the continuation of the current extra-budgetary post in the UNECE secretariat, through additional donor funding. As above, if this is not forthcoming, either the timetable will need to be extended, or the outputs will need to be re-defined.

20. All four work packages will run throughout the year, though substantial work should be completed by mid-November so that outcomes can be reported and demonstrated at the HLG Workshop.

**VIII. Project governance**

21. 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.

22. 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.

1. *What does Big Data mean for Official Statistics?* available at <http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170622> [↑](#footnote-ref-1)
2. *Strategy to Implement the Vision of the HLG* available at <http://www1.unece.org/stat/platform/display/hlgbas/HLG+Strategy> [↑](#footnote-ref-2)