Statistical Journal of the IAOS 31 (2015) 159–163
DOI 10.3233/SJI-150889
IOS Press

International collaboration to understand the relevance of Big Data for official statistics

Steven Vale
United Nations Economic Commission for Europe, Palais des Nations, CH-1211, Geneva, Swizerland
Tel.: +41 22 917 3285; E-mail: steven.vale@unece.org

Abstract. This paper describes international collaboration activities to help the official statistical community to better understand the phenomenon of "Big Data", and how it might impact on statistical production. This work is taking the form of annual projects, overseen by the High-Level Group for the Modernisation of Statistical Production and Services, a group of ten heads of national and international statistical organisations. The results obtained during 2014 are presented, and the paper looks forward to the planned activities in 2015.

Keywords: Big Data, international collaboration, sandbox

1. Background

At a High-Level Seminar on Streamlining Statistical Production and Services, held in St Petersburg, 3–5 October 2012, participants asked for "a document explaining the issues surrounding the use of big data in the official statistics community". They wanted the document to have a strategic focus, aimed at heads and senior managers of statistical organisations.

To address this requirement, the High-Level Group for the Modernisation of Statistical Production and Services (HLG) [1] established an informal Task Team, which prepared the paper "What does 'Big Data' Mean for Official Statistics" [2]. Further discussions on this topic during 2013 resulted in a proposal for a major international collaboration project under the HLG. This project was approved in November 2013 at an annual workshop on statistical modernisation, which brings together chief statisticians and representatives of various expert groups and projects working in the area of statistical modernisation.

The project ran during 2014, and the outputs are described in detail in the following sections. It had three main objectives:

- To identify, examine and provide guidance for statistical organizations on 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.

The project brought together around 75 people from 25 national and international statistical organisations. The infographic in Fig. 1 shows the participation in, and the structure of, the different project activities.

As the project was judged to be successful, the November 2014 modernisation workshop decided to support a follow-up project in 2015, building on the previous results. The challenges for 2015 are outlined towards the end of the paper.

The projects described in this paper are designed to enhance knowledge and skills within official statistics to meet the challenges of new data sources. In doing so, they provide a firm basis for further work such as the projects recently launched by the United Nations Statistical Division and Eurostat. Both of these organisations have contributed actively to the work described below. These projects should be seen as one element

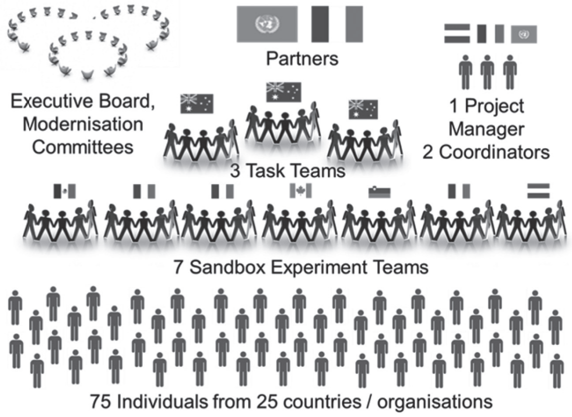
Partners
Executive Board, Modernisation Committees
1 Project Manager 2 Coordinators
3 Task Teams
7 Sandbox Experiment Teams
75 Individuals from 25 countries / organisations

Fig. 1. Big data project participants and structures.

of the wider initiative to modernise the production of official statistics, under the HLG and the Conference of European Statisticians.

2. Establishing the priorities

All international collaboration projects under the HLG, facilitated by the UNECE, are demand driven. Participation is voluntary and unpaid, so the topics treated are, by definition, of value to the participating organisation. If a topic is proposed but there are insufficient volunteers to undertake the necessary work, the topic is dropped or merged with another. For this reason, it was essential early in the project to determine the priorities of the participating organisations.

Following a call for expressions of interest to statistical organisations worldwide, a group of volunteers was convened for a "virtual sprint" session (for more information on the nature and use of "sprints", please see Lalor and Vale [3]). A number of issues were identified, and were quickly summarised in a discussion paper "How Big is Big Data?" which was circulated to the wider community for feedback. The results were then discussed at a face-to-face "sprint" in Rome, which set the following priority areas for the project:

- Quality
- Partnerships
- Privacy
- Skills
- Methodology and technology

Task Teams were established to consider and produce guidelines on the issues around quality, partnerships and privacy. The existing Modernisation Committee on Organisational Frameworks and Evaluation was asked to conduct a survey on the skills needed to work with Big Data, and the training requirements to deliver those skills. Finally, methodological and technological issues were passed to the "Sandbox" Task Team.

3. Introducing the "Sandbox"

A web-accessible environment for the storage and analysis of large-scale datasets was created with support from the Central Statistics Office of Ireland and hosted by the Irish Centre for High-End Computing (ICHEC). This environment, known as the "Sandbox" provides a computing environment to load Big Data sets and tools, with the goal of exploring how they could be used for statistical production. It was launched in April 2014 at a workshop in Dublin, and gives participating statistical organisations the opportunity to:

- Test the feasibility of remote access and processing – Statistical organisations around the world can access and analyse Big Data sets held on a central server.
- Test whether existing statistical standards/models/methods etc. can be applied to Big Data;
- Determine which Big Data software tools are most useful for statistical organisations;
- Learn more about the potential uses, advantages and disadvantages of Big Data sets – "learning by doing";
- Build an international collaboration community to share ideas and experiences on the technical aspects of using Big Data.

The Sandbox infrastructure is a shared distributed computational environment composed of 28 machines ("nodes") which are physically located within the ICHEC data centre and are connected to each other through a dedicated, high-speed network. It contains the following Big Data software tools:

- Hadoop (HDFS and MapReduce) – open-source software project, developed by the Apache Software Foundation, which splits large data sets into smaller chunks for parallel processing
- Pig and Hive – Pig is a high level interface to MapReduce, based on a high-level language, "PigLatin". Hive is an interface to MapReduce that allows data to be structured in tables, as in a relational database.

- RHadoop – a tool to write MapReduce programs in the programming language “R”, which is familiar to many statisticians.
- Spark – a distributed computation tool that exploits in-memory computation to accelerate processing operations on distributed datasets.
- Pentaho – visual analytics software to facilitate understanding of data (including distributions, outliers etc.) especially when applied to Big Data sources.

Several Big Data datasets were acquired and loaded into the sandbox environment. Seven experiment teams were established, typically comprising between 4 and 6 methodologists and IT experts from different countries. These teams covered the following topics:

- Consumer price indices – experimenting with the computation of price indexes using the different tools available in the Sandbox, and synthetic data sets modelling scanner data created using software developed by the team.
- Mobile telephone data – exploring the possibility of using data from mobile telephones as a source for computing statistics on tourism, daily commuting etc. The team used real data in aggregated form acquired from the telecom provider Orange.
- Smart meters – computing statistics on power consumption using data collected from smart meter readings. Two data sets were used: data from Ireland and a synthetic data set from Canada.
- Traffic loops – creating traffic statistics using data from traffic loops installed on roads in the Netherlands.
- Social media – using Twitter data from Mexico to analyse sentiment and to tourism flows.
- Job portals – computing statistics on job vacancies starting from job advertisements published on web portals.
- Web scraping. This team tested different approaches for automatically collecting data from web sources.

4. Project results

The different teams working on the 2014 Big Data project produced the following results by the end of 2014. A set of deliverables describing these results in detail was released on the UNECE Big Data Wiki [4].

4.1. Quality

The main output of the Quality Task Team was a framework for the quality of Big Data. The team found that:

- There is a need for quality assessment covering the entire business process;
- Input quality can be explored and assessed by using and elaborating existing input quality frameworks;
- Throughput quality can be maintained by following quality processing principles, but throughput quality dimensions need to be further developed for Big Data processing;
- Additions have been proposed to output quality dimensions from existing frameworks, to make them suitable for Big Data applications.

4.2. Partnerships

This Task Team used information from surveys of national and international statistical organizations, conducted in collaboration with the United Nations Statistical Division. Their main output was a set of guidelines for the establishment and use of partnerships in Big Data projects for official statistics. They found that:

- Most partnership arrangements encounter similar issues related to financial/contractual arrangements, legislative, privacy and confidentiality issues, responsibilities and ownership issues and other risks;
- The importance of these issues depends on the type of partner;
- For the Sandbox experiments, the main issue was timely access to data;
- A project can only exist if a working partnership can be forged with a data provider to provide a reliable data source.

4.3. Privacy

The team produced papers taking stock of the current status of statistical disclosure control, investigating the characteristics of Big Data and their implications for data privacy, as well as a summary of practical measures to manage Big Data privacy. They found that:

- Existing tools are well-developed;
- Privacy risks can be linked to Big Data characteristics;
- But: There is not much experience yet with Big Data privacy issues.

4.4. Skills

The key findings of the survey on skills for working with Big Data in statistical organisations were:

- 37% of respondents already work with Big Data, and 43% are planning to work with Big Data in the near future;
- The 3 most important types of skills for working with Big Data were identified as:
  - ∗ IT skills: noSQL databses, SQL databases and Hadoop
  - ∗ Statistical skills: Methodology and standards for processing Big Data, data mining
  - ∗ Other skills: Creative problem solving, data governance and ethics
- At present there is insufficient training in these skills

4.5. Methodology and technology

Some of the key findings of the Sandbox Task Team were:

The Sandbox Task Team found that:

- A common computing environment enables shared work on methodology, especially where the data sets have the same form in all countries. Methods can be developed and tested in the shared environment and then applied to real data sets within each organization;
- Although web sources and social media are appropriate for international sharing, language differences can present a problem when cleaning and classifying text data;
- Other work on methodology was done in the mobile phones team, for computing the movement of people starting from call traces, and in the traffic loops team, for calculating the average number of vehicles in transit for each day and for each road;
- The project shows for the first time, on a practical basis and on a broad scale, the potential and the limits of using Big Data sources for computing statistics. Improvements in efficiency and quality are possible by replacing current data sources. New products can be obtained from novel sources such as traffic loops, mobile phones and social media data. However, some sources can be of low quality and require some serious pre-processing before use;
- The use of a shared environment for the production of statistics is severely limited by privacy constraints. These limitations can be partly bypassed through the use of synthetic data sets. Another solution is to generate the data by modelling its behaviour. Both approaches were used in the project, for smart meters data and scanner data respectively;
- Big Data tools are essential when the size of data sets is measured in terms of hundreds of gigabytes or larger. They can be more efficient than existing data processing tools for data larger than one gigabyte;
- Researchers/technicians should be able to master different tools and be ready to deal with immature software, so strong IT skills are needed.

5. Outlook for 2015

At the time of writing (January 2015), a follow-up project is in the start-up phase. This project will build on the success of the previous one, and will ensure that some of the networks of expertise established in 2014 can continue to deliver benefits for the official statistics community.

The aims of the 2015 project are:

- To extend the range and scope of experiments using the Sandbox;
- To continue determining and developing the skills needed to work with Big Data within the official statistics community;
- To enhance the existing inventory of Big Data projects and activities within statistical organisations.

In addition to the above, the HLG has set the project team the challenge to develop and publish a set of internationally comparable statistics based on one or more Big Data sources by the end of 2015.

6. Conclusions

There is a lot of hype around the topic of Big Data. This can lead to either unrealistic expectations of what Big Data might deliver, or an assumption that it is just a passing trend and can be ignored. One thing that is becoming increasingly clear to many working in official statistics, is that the rapid growth in new data sources (whether they are “big” or not) is likely to have a profound impact on official statistics, in the same way that the use of administrative sources has changed the way statistics are produced in many countries.

Each statistical organisation faces the same challenges to understand the potential benefits and constraints of the new data sources. It is therefore logical to work together within the official statistics "industry" to share ideas and experiences, so that we can move more rapidly and efficiently to a position where new sources can contribute effectively to the production of both traditional and new statistics.

References

[1] UNECE High-Level Group Wiki, http://www1.unece.org/stat/platform/display/hlgbas.

[2] UNECE, What Does Big Data Mean for Official Statistics, 2013, http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614.

[3] T Lalor and S Vale, Sprints – Lessons Learned from HLG Projects, 2013, http://www1.unece.org/stat/platform/display/hlgbas/Sprints+-+Lessons+learned+from+HLG+projects.

[4] UNECE Big Data Wiki, http://www1.unece.org/stat/platform/display/bigdata.