Project Overview Report

1. Project Introduction

The strategy of the High-Level Group for the Modernisation of Statistical Production and Services (HLG) 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". This project is aligned with these aspirations since it focuses on new sources, new methods, new outputs, and ways to tackle the issues surrounding these.

The overall aim of the project was to 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. Its purpose was to tackle strategic and practical issues that are multi-national in nature, rather than those that are specific to individual organizations or national sources. While the project involved a practical component and a consideration of methodological issues, its aim was 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.

The project had three main objectives:

1. 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.
2. 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.
3. To facilitate the sharing across organizations of knowledge, expertise, tools and methods for the production of statistics using Big Data sources.

According to the original project proposal, this project will be successful if it resulted 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 were:

- Work Package 1 - Issues and Methodology: a consistent international view of Big Data opportunities, challenges and solutions, documented and released through a public web site.
- Work Package 2 - Shared computing environment ('sandbox') and practical application: 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 - Training and dissemination: exchange of knowledge and ideas between interested organizations and a set of standard training materials.
- Work Package 4 - Project management and coordination: the project is completed on schedule, and delivers results that are of value to the international statistical community.

The project provided the expected results in all four work packages, and delivered more than initially expected in the Sandbox where we succeeded to acquire, load and analyse seven different types of datasets including a pre-production test of full Netherlands Road Sensor dataset. Due to the significantly longer time required for data acquisition than expected, we could not, however, perform all the analyses that we would like to, so we decided to ask Executive Board for extension of Sandbox environment, which was approved. Most existing experiments will therefore continue until March 2015, and will provide valuable findings for Big Data project 2015, which has the ambitious goal to develop the first multi-national official statistics, based on one or more global Big Data sources.

The project was composed of four task teams:

- Quality Task Team, with 13 members from 9 national and international statistical organizations. They were responsible for designing a complete quality framework for Big Data.
- Privacy Task Team, with 11 members from 9 national and international statistical organizations and academic / research institutions. Their objective was to review existing tools for risk management in view of privacy issues, focusing on Big Data characteristics and their implications for data privacy.
- Partnership Task Team, with 12 members from 8 national and international statistical organizations. They were tasked to explore partnership models with organizations such as academia, scientific communities, research institutes, and technology providers.
- Sandbox Task Team, with 38 participants from 18 national and international statistical organizations. Their aim was to design, install and use a web-accessible computing environment where researchers coming from different institutions could explore tools and methods needed for statistical production verifying in practice the feasibility of producing Big Data-derived statistics. The team was organized around a set of "experiment activity teams", focusing on topics related to different statistical domains.

The results could not have been achieved without the excellent cooperation with many organizations and companies. The Central Statistics Office in Ireland arranged the Big Data computing environment and national statistical organizations in Italy and the Netherlands hosted face to face sprints and workshops. In the Sandbox there was an excellent cooperation with technology providers for example ICHEC and Hortonworks. Data was provided or facilitated by several national statistical organizations: Canada, Ireland, Italy, Mexico, the Netherlands and mobile provider Orange. The work of Partnerships task team was underpinned by the information from questionnaire developed jointly with the United Nations Statistical Division, in preparation for the International Conference on the Use of Big Data for Official Statistics, which was held in Beijing, China, in October 2014.

2. Issues and Methodology

2.1. Introduction

During the initial phase of the project, a team of experts met for a virtual sprint session, followed by a workshop in Rome, and identified six main priorities for investigation during the project:
The first four topics were assigned to task teams within the Big Data project. The remaining two were included in the scope of the work of Modernisation Committees. The resulting deliverables - frameworks, guidelines and recommendations are based on the work of dedicated teams of experts but also on experience and practical work in the Sandbox experiments. Results for first three priorities are described in this chapter. Findings for methodology, technology and skills are described in the following chapter.

2.2. Quality

Introduction

The Quality Task Team assessed several existing quality frameworks for official statistics with respect to their applicability to Big Data (frameworks such as those produced by Statistics Sweden, Statistics Canada, the Australian Bureau of Statistics, the Statistical Network, and the European Statistical System Code of Practice).

There are commonalities across official statistics quality frameworks that are worth noting. Firstly there is a widespread adoption of a dimensional approach. Quality is segregated into a number of dimensions and each dimension is assessed independently of the others. Commonly found dimensions include accuracy, coherence, relevance, clarity, and timeliness, as well as others. This dimension approach arises from the total quality approach, which holds that when assessing quality, all aspects of a product that are relevant to the user must be considered.

While existing frameworks provide a well-rounded approach to statistical data quality, they primarily developed for use with survey data collection. There are implications for the use of Big Data that may not be captured by these frameworks in the receipt, processing, and reporting of statistical products.

Furthermore, existing frameworks tend to be output focused. This reflects the high degree of control that statistical organizations have previously had over the creation and initial processing of data used in statistical products. In a Big Data context, there may be provenance or other quality issues around the data source that are not typically considered in traditional data collection contexts.

The concept of ‘fitness for purpose’, the notion that the quality of data can only be evaluated in light of its intended use, is a central idea in data quality, and remains so in the context Big Data.

Big Data Quality Framework

The team concluded that extensions to existing statistical data quality frameworks were needed in order to encompass the quality of Big Data. A preliminary framework was developed building on dimensions and concepts from existing statistical data quality frameworks. The Big Data Quality framework as developed by the UNECE Task Team provides a structured view of quality at three phases of the business process:

- **input** – when the data is acquired, or in the process of being acquired
- **throughput** – any point in the business process in which data is transformed, analysed or manipulated
- **output** – the reporting of quality with statistical outputs derived from Big Data sources.

The framework uses a hierarchical structure composed of three hyperdimensions with quality dimensions nested within each hyperdimension. The three hyperdimensions are the source, the metadata and the data. The concept of hyperdimensions has been borrowed from the administrative data quality framework developed by Statistics Netherlands. (See Struijs et al, 2013);

The hyperdimension "source" relates to factors associated with the type of data, the characteristics of the entity from which the data is obtained, and the governance under which it is administered and regulated.

The hyperdimension "metadata" refers to information available to describe the concepts, the contents of the file, and the processes.

The hyperdimension "data" relates to the quality of the data itself.

At the input phase of the business process, a statistical organization should engage in a detailed quality evaluation of a Big Data source both before acquiring the data (this known as the ‘discovery’ component of the input phase), and after (this is the ‘acquisition component’).

In addition to dimensions commonly used to assess statistical output, the task team recommended the use of new dimensions, including privacy and confidentiality (a thorough assessment of whether the data meets privacy requirements of the organization), complexity (the degree to which the data is hierarchical, nested, and comprises multiple standards), completeness (of metadata) and linkability (the ease with which the data can be linked with other data).

For the throughput phase of the business process, four principles of processing are proposed:

1. System Independence: The result of processing the data should be independent of the hardware and software systems used to process it;
2. Application of quality dimensions: in monitoring process quality, quality dimensions as articulated here and in other frameworks should be used as a guide for quality evaluation;
3. Steady states: that the data be processed through a series of stable versions that can be referenced by future processes and by multiple parts of the organization;
4. Application of Quality gates: that the organization should employ quality gates as a quality control business process.

For the output phase of the business process, the team used the Australian Bureau of Statistics Data Quality Framework as a starting point. Two new output dimensions were recommended for output quality: confidentiality and complexity. In addition, two subdimensions were proposed for the dimension of coherence: consistency and validity.
The input component of the framework was tested by the Job Vacancies project component of the Sandbox team. The framework was found to be a useful tool for analysing the data and in particular illuminated the data source in two notable ways: it highlighted the high quality of the job vacancy data in terms of timeliness, but also drew attention to the difficulty of assessing the coverage.

Conclusion

In conclusion, rather than focusing quality efforts only on statistical outputs of Big Data, statistical organizations need a series of quality frameworks across the business process. The UNECE Quality Task Team has recommended some principles as well as dimensions that would be useful for evaluating Big Data sources and products.

2.4. Privacy

Introduction

The Privacy Task Team was asked to give an overview of existing tools for risk management in view of privacy issues, to describe how risk of identification relates to Big Data characteristics, and to draft recommendations for statistical organizations on the management of privacy risks related to the use of Big Data. The Privacy Task Team was comprised of representatives from several statistical organizations as well as academics and researchers from all over the world.

The Task Team concluded that extensions to existing frameworks were needed in order to deal with privacy risks related to the use of Big Data. This report summarises the outcome of the Big Data Privacy Task Team.

Existing tools for privacy risk management

Statistical organizations try to manage two conflicting aims: providing access to data for the benefit of society, and to meet society’s expectation that sensitive information about data providers will be kept private. For traditional data sources, in particular sample surveys and administrative data collected by governments, the Task Team has considered what types of risk exist. Disclosure can occur for estimate releases, as well as for micro-data releases, except that information about a unit must first be derived from one or more estimates. An extensive body of literature on this is available. The risk of disclosure is also influenced by the type of access.

As to the management of disclosure risk, two broad areas are:

1. Reducing the risk that an attempt at disclosure will be made
2. Releasing data in a manner that it is not likely to enable disclosure

For each of these areas, a number of factors can be identified that affect the risk of disclosure.

Within this framework, the Task Team has specifically looked into the different ways in which statistical organizations allow analysts or researchers to access micro-data, managing the risk of disclosure associated with databases, and the advantages and disadvantages of the different approaches to managing privacy. Relevant software was also identified.

As to micro-data access, statistical organizations have developed different strategies to enable external researchers to analyse their data without violating confidentiality regulations. These strategies fall under three broad topics: micro-data dissemination, onsite analysis in research data centres, and remote access.

Concerning risks related to databases, it is useful to make a distinction between owner privacy, respondent privacy and user privacy. Owner privacy refers to situations where entities make queries across their databases in such a way that only the results of the query are revealed. Respondent privacy is about preventing re-identification of respondents, e.g. individuals or organizations. User privacy is about guaranteeing the privacy of queries to interactive databases, which is necessary to prevent user profiling and re-identification.

As to the advantages and disadvantages of approaches to managing privacy, one has to bear in mind that when an organization applies statistical disclosure control to micro-data, or estimates derived from them, there is an implicit trade-off between the disclosure risk and the utility. Disclosure risk depends upon the context in which the attack is assumed to occur. While a wide range of attacks are possible, a pragmatic approach for an agency to take is to focus on the scenarios that are likely to have the greatest disclosure risk. The Task Team has looked into disclosure risk and utility for queryable databases, micro-data and tabular data.

Big Data characteristics and privacy risk

Big Data is often characterised by three or four Vs: velocity, variety, veracity, and sometimes also value. However, there are more aspects that may be relevant to privacy risk. In particular size, availability, aggregation, awareness of society, flexibility, provider infrastructure and geographical differences may be pertinent.

The existing tools for privacy risk management can be assessed for their application to Big Data by taking these characteristics into account. For instance, of the three micro-data access options (i.e. micro-data dissemination, onsite analysis in research data centres, and remote access), micro-data dissemination is no longer an option in most cases, since one feature of Big Data is that the size of the data implies that transferring the data is cumbersome. It would not be possible to send the data to the researcher on compact discs, or provide a link for a download as is the current practice with micro-data dissemination. Thus, onsite and remote access will be the only viable solutions.

Onsite access has the advantage of providing the agency with better control over who accesses the data and what can be done with the data. However, if the standard set of surveys and administrative data that is already offered by the statistical organization will be enriched by new data sources it is likely that offering manual output checking for all these databases will no longer be feasible.
As to the release of estimates, the common practice of re-identification experiments will probably not be useful in the Big Data context. Disclosure risks are typically assessed by assuming that the intruder has information on some of the variables in the dataset from another data source and then tries to use this information to identify somebody in the confidential dataset. Under this scenario record linkage experiments are run to see how many records would be matched correctly. Given the fact that such experiments usually have already long running times with survey data, their feasibility to Big Data is questionable. It is thus more relevant than ever that general strategies for ensuring the confidentiality of the generated outputs without manual interference are developed.

Recommendations

For a Big Data context, a number of guidelines for risk treatment can be given, some of which build on existing tools, whereas others are more novel. The recommendations given by the Task Team can be grouped into:

1. Information integration and governance:
   a. Database activity monitoring, i.e., keeping track of who has access to your databases and what they are doing at any given time.
      i. Application of best practices for security of IT systems and business practices. As a baseline, four best practices should be applied: separation of duties, separation of concerns (a modular approach to functionality where possible), the principle of least privilege (no more access rights than needed), and defence in depth (multiple security mechanisms).
      ii. Separation of Duties;
   b. Separation of Concerns (a modular approach to functionality where possible);
      i. Principle of Least Privilege (no more access rights than needed); and
      ii. Defence in Depth (multiple security mechanisms/layers).
   c. Application of best practices of security of transportation, such as Transport Layer Security (TLS).
   d. Data encryption. Examples are Full Disk Encryption (FDE) and File System-level Encryption (FSE).

2. Statistical disclosure limitation/control (identity disclosure, attribute disclosure, inferential disclosure and population/model disclosure):
   a. Preserving confidentiality by restricting data access and restricting data release.
   b. Ensuring access to useful data. It is important to ensure that the released data, while not identifiable, fulfils a purpose and is still meaningful [3].
   c. Balance data utility and disclosure risk. Apart from traditional approaches, modern techniques are proposed, such as the use of statistical models to simulate data records that emulate the main distributional characteristics of the original micro-data.

3. Managing potential risk to reputation (public image):
   a. Enforce ethical principles in the supply chain, including legal and administrative instruments of accountability and informed consent.
   b. Establish strong compliance controls.
   c. Develop a monitoring system to track reputational threats.
   d. Ensure transparency and understanding through clear communication with stakeholders, for instance on the use of data on citizens, and organisation of a dialogue with the public.
   e. Create a crisis communication plan.

2.5. Partnerships

Introduction

The need to collaborate with other organizations is identified as one area of strategic importance. The reality is that no statistical organization alone can take advantage of the opportunities, or respond to the challenges, that Big Data brings; even together the industry would struggle to develop the access to data sources, analytical capability and technology infrastructure needed to deliver Big Data strategies.

Partnership with data providers and sources is an important and often first step, but in order to optimise the use of the data, other types of partnerships might be needed. Statistical organizations might need to collaborate with each other but also to partner with academics, scientific communities, research institutes, and technology providers, not only to develop Big Data standards, processes and methodologies, but also to gain access to analytical capability and access to the most advanced technology. Relationships with other stakeholders such as those concerned with ethical and privacy issues may also be key in order to build the trust and support required for any statistical organization.

A task team was set up with representatives from statistical organizations across the world to examine the known issues around partnering with different types of organization within a Big Data context. General guidelines were prepared on establishing partnerships for Big Data.

Methods

Partnership opportunities need careful consideration across various themes to ensure that an optimal arrangement is agreed, for both statistical organization(s) and the partner(s) involved. In order to better understand the themes to be considered, and the approaches necessary to set up a successful partnership with respect to Big Data it is necessary to first obtain information on the current experiences across a range of statistical organizations.

Two questionnaires were sent to statistical organizations in a joint exercise with the United Nations Statistical Division. The first focused on the overall strategy for Big Data in the organization, while the second focused on specific projects on Big Data. The questionnaire identified a number of different partnerships arrangements that contributed to the development of a report containing guidelines to consider when partnering on Big Data projects.

Feedback from the other task teams within the UNECE project was also incorporated into the report. These teams specifically provided input on the Privacy and Quality dimensions of partnering on Big Data projects but also the partnering arrangements experienced within each of the individual pilot projects in the practical Sandbox element of UNECE’s overall project.

Results
The examples from the questionnaire and Sandbox demonstrated how most partnership arrangements encounter similar forms of issues related to financial /contractual arrangements, legislative, privacy and confidentiality issues, responsibilities and ownership issues and other risks to be managed. Of course, the relative importance of these issues depends on the type of partner, be it partnerships with data providers or design, technology, IT infrastructure or analysis partners. However, many of the issues are comparable, and drawing upon the experiences from statistical organizations, the report of the task team highlights some important aspects to facilitate future partnerships on Big Data projects.

The questionnaire indicated that some organizations have been tapping and expanding the use of administrative data owned by line ministries and agencies (e.g., taxation records, civil registry, customs records, etc.). Other data providers are from the private sector (e.g., internet companies, mobile phone operators, etc.) and many of these players are multinational enterprises. Therefore, it is important to establish good relationships with both government and private sector partners.

At the national level, some projects are already completed, but many more are in the early stages of implementation or at the idea stage. Most of these projects involve at least one partner, often from the private sector or academia. However, the questionnaire has identified that most partnerships are arranged individually and sometimes on an ad-hoc basis, with some partnerships more successful than others. There are therefore important lessons learned and experiences from different organizations that should be shared in order to facilitate better and more fruitful Big Data partnerships in the future.

Well-functioning partnerships with data providers are seen as the crucial aspect in the successful implementation of any Big Data project. In the sandbox experiments the biggest challenge was timely access to data, often due to concerns of confidentiality. In many cases a project can only exist if a working partnership can be forged with a data provider. While the technical issues of data analysis are certainly challenging, they are often secondary to establishing a reliable data source. Furthermore, since data providers often aggregate and clean data, a good working relationship ensures clear communication of methodology and, at times, allows for much of the data cleaning to be completed by the provider.

Conclusions

This task team has provided a number of initial guidelines for partnerships in Big Data related projects. By building on experiences and lessons learned in a number of countries, specific guidelines for financial/contractual arrangements, legislative, privacy and confidentiality issues, and responsibilities and ownership have been established. However, further work is needed to provide more detailed and operational guidelines that could perhaps serve as a framework for partnership arrangement in Big Data projects. Such a partnership framework could include templates for different partnership agreements. These templates would provide standard suggestions for the various issues that need to be defined in each partnership agreement, such as financial issues, privacy and confidentiality, and other issues.

3. The Sandbox

3.1. Introduction

The general purpose of the Sandbox work package was to evaluate the feasibility of a shared approach for using Big Data sources and tools for producing official statistics, through collaborative work in a shared computation environment.

The Project Proposal defines 5 specific objectives for stating successful completion of the work:

1. 'Big Data' sources can be obtained (in a stable and reliably replicable way), installed and manipulated with relative ease and efficiency on the chosen platform, within technological and financial constraints that are realistic reflections of the situation of national statistical offices
2. The chosen sources can be processed to produce statistics (either mainstream or novel) which conform to appropriate quality criteria – both existing and new – used to assess official statistics, and which are reliable and comparable across countries
3. The resulting statistics correspond in a systematic and predictable way with existing mainstream products, such as price statistics, household budget indicators, etc.
4. The chosen platforms, tools, methods and datasets can be used in similar ways to produce analogous statistics in different countries
5. The different participating countries can share tools, methods, datasets and results efficiently, operating on the principles established in the Common Statistical Production Architecture.

In this chapter we first provide an overview of the Sandbox environment, presenting the available tools. Then, we describe the experiments that have been carried out using the Sandbox, as well as the data sources that were acquired. Findings from each experiment are illustrated, highlighting if and how experiments meet each of the 5 specific objectives (referred to as SO1 to SO5). Finally, a section is devoted to presenting the results concerning technology topics, stemming from the experience of the work group with the new tools.

The Sandbox Environment

A web-accessible environment for the storage and analysis of large-scale datasets has been created and used as a platform for collaboration across participating organizations. The “Sandbox” environment was created, with support from the Central Statistics Office (CSO) of Ireland and the Irish Centre for High-End Computing (ICHEC). It provides a technical platform to load Big Data sets and tools, 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.

The Sandbox infrastructure is a shared distributed computational environment composed of 28 machines running a Linux operating system. The nodes are physically located within the ICHEC data center and are connected with each other through a dedicated, high-speed network.

Tools in the Sandbox

Hadoop (HDFS and MapReduce)
Hadoop is an open-source software project, developed by the Apache Software Foundation, targeted at supporting the execution of data-oriented applications on clusters of generic hardware. The core component of the Hadoop architecture is a distributed file system implementation, namely HDFS. HDFS realizes a file system abstraction on top of a cluster of machines. Its main feature is the ability to scale to a virtually unlimited storage capacity by simply adding new machines to the cluster at any time. The MapReduce paradigm is a programming model specifically designed for the purpose of writing programs whose execution can be parallelized. A MapReduce program is composed by two functions, "map", specifying a criteria to partition the input into categories and "reduce" where the actual processing is performed on all the input records that belong to a same category. The distributed computation framework (e.g. Hadoop) takes care of splitting the input and assigning it to a different node in the cluster that takes in turn the role of "mapper" and "reducer". Hence, the general processing task is split into separate sub-tasks and the result of the computation of each node is independent from all the others. The physical distribution of data through the cluster is transparently handled by the platform and the programmer must only write the processing algorithm according to the MapReduce rules. MapReduce is a general paradigm not tied to a specific programming language. While Hadoop requires MapReduce programs to be written in Java, MapReduce interfaces exists for all the common programming and data-oriented languages, including R.

**Pig and Hive**

Pig is a high level interface to MapReduce. Writing MapReduce programs in Java can be complex and also common operations on data, like joins and aggregations, may require a significant amount of code. This requires trained IT developers, slowing down the analysis of data. Pig is based on a high-level language, namely PigLatin, oriented to data transformation and aggregations. Complex operations on data can be performed with short scripts, that can be simply read and modified also by business analysts.

Hive is a SQL interface to MapReduce. It allows to structure data in HDFS in tables, like in a relational database, and to query data using familiar SQL constructs such as selections, joins and filters. Hive and Pig have different purposes and complement themselves in the Hadoop ecosystem: Hive is oriented to interactive querying of data, while Pig allows to build complex transformation flows.

**RHadoop**

RHadoop is a R library that allows to write MapReduce programs in R. Once installed and configured, it integrates with the Hadoop cluster, allowing to read and write files from/to HDFS and to stream MapReduce jobs over the cluster. The advantage it provides consists in the possibility using a language which is already familiar to statistical users, allowing them to work on Big Data with a limited learning curve and exploiting established know-how. It is important to point out that RHadoop is a purely open-source project which currently is not backed up by any company. This means that only a few contributors actively develop the software, so that one cannot expect commercial-level software maturity for aspects such as stability and documentation. However, the project is followed by an active community that can help dealing with issues.

**Spark**

Spark is a novel distributed computation infrastructure that promises to overcome the performance limitations of MapReduce. The idea behind Spark is to exploit in-memory computation for accelerating operations on distributed datasets. This is different to MapReduce, which is mainly based on input-output disk operation in all the phases of the computation. Spark is built on the concept of Resilient Distributed Dataset, a collection of data distributed on different machines on which transformations can be applied. Basically, from a programmer's point of view, Spark is an API available for three different languages (Scala, Python and Java) that allows to handle RDDs. Spark can be integrated with an Hadoop distribution so that it can read and write files from/to HDFS.

In contrast with MapReduce, Spark does not yet have high-level interfaces like Pig and Hive, so working with Spark involves learning Python, Scala or Java. An integration with R in the form of a package (SparkR) is already available but currently appears to be in an early state of release.

**Pentaho**

Visual analytics software is a class of tools that allow to quickly build visual representations of data that allow for interactive discovery and fast plotting. Visual data discovery is an effective technique for deriving global characteristics of data (such as distributions, outliers etc.), especially when applied to Big Data sources.

Pentaho Enterprise platform is a visual analytics tool that can connect to several data sources (databases, text files etc.) and let the users build graphs and visualizations of data. Different visualizations can be assembled in dashboards through which data can be navigated interactively. Pentaho is an open source project with a commercial enterprise edition that incorporates full functionality. We were able to obtain a trial license of the enterprise edition extended through the project lifespan through a partnership with Italian distributor BNova.

We tested the connection of Pentaho to data in the Sandbox through Hive but we found that the high latency of the response was not suited to perform analysis. Then we decided to load data in a open-source database shipped with Pentaho, namely MonetDB. MonetDB is a columnar-store database, a kind of technology largely used for backing analytics applications because it provides better performance with respect to relational databases when querying millions of rows with a limited number of columns.

### 3.2. The Experiments

The experiment teams, along with a short description of their objectives:

- **Consumer Price Indexes.** Worked on testing performance of Big Data tools by experimenting with the computation of price indexes through the different tools available in the Sandbox. The data sources are synthetic data sets that model price data recorded by point of sales in supermarket (“scanner data”), the use of which within price statistics is currently under study in several statistical organizations. The main contribution of this task team is an assessment of the performance of Big Data tools, applied to synthetic data sets modelling scanner data. Synthetic data was generated through dedicated software developed as part of the work of the team. Following this approach, we were able to easily test data sets of various sizes, working with both Big Data tools and “traditional” software such as statistical tools and relational databases, all available in the Sandbox environment.

- **Mobile Phone Data.** Worked on exploring the possibility of using data from telecom providers as a source for computing statistics on tourism, daily commuting etc. The team used real data in aggregated form acquired from the French telecom provider Orange.

- **Smart Meters.** Experimented with the computation of statistics on power consumptions starting from data collected from smart meter readings. Two data sets are available: data from Ireland and a synthetic data set from Canada.
• Traffic Loops. Worked on computing traffic statistics using data from traffic loops installed on roads in the Netherlands. Methodological challenges for cleaning and aggregating data were addressed. Moreover, the team will use the Sandbox environment for testing the possibility of computing production statistics over a huge dataset (6Tb).

• Social Data. This team explored the possibility of using Twitter data generated from Mexico and collected along several months for analyzing sentiment and to detect touristic activity. A collection of Mexican tweets was studied in relation to their sentiment. Because Mexican tweets can be written in different languages, such as Spanish, English and some local dialects, first studies focused on the applicability of smileys and emotions as an indicator for the sentiment in public messages. Twitter messages produced in the first 5 months of 2014 were studied.

• Job portals. Worked on computing statistics on job vacancies starting from job advertisements published on web portals. Tested the approach by collecting data from portals in various countries. After a first exploration phase where data from various countries was considered, the team set up an automated daily collection of data from 7 Italian web portals.

• Web scraping. This team tested different approaches for collecting data from web sources.

Data acquisition

• 7 different datasets were loaded in the Sandbox environment. Acquisition of the datasets has been a complex task from which several lessons have been learned. One obvious outcome was that data sets that are more “interesting” from a statistical point of view, that is carrying more information and not aggregated, are in general more difficult to retrieve, since they are limited by privacy constraints.

• The mobile phone dataset from Orange was an interesting case in this sense. It represents data from Ivory Coast and it was used for a competition promoted by the company, where researchers were invited to use the data for experiments on Big Data. Despite the fact that the data was already freely released to researchers in the context of the competition, we had to go through a process of legal review, only because the purpose of utilization was different from the original one. Smart meter data from Ireland was released under similar conditions.

• In order to adhere to the terms and conditions of the above mentioned data sets we had to enforce stricter privacy in part of the Sandbox environment, implementing access control in order to authorize access to sensitive data only for users in the relevant team.

• Sources from the web and social media were also used in the Sandbox experiments, in particular data from Twitter and from job portals in Slovenia and Italy, as well as price data from e-commerce web sites. Although this form of data is easily available, we experienced issues with quality, in terms of coverage and representativeness. We tested different techniques and tools for scraping data.

• Finally, we cite the case of some data sets that could not be acquired. Satellite data from Australia, to be used for agricultural statistics, could not be released by the providers in time for the first phase of the project and is planned to be loaded in a possible extension of the project. There may also be possibilities to acquire other interesting data sets, such as marine and air transport data.

Summary of Findings

• A common computation environment enables shared work on methodology, especially where the data sets have the same form in all countries, so methods can be developed and tested in the shared environment and then applied to real counterparts within each statistical organization. Examples of these data sets are smart meters, scanner data, web sources and social media. As examples of this approach we tested the application of a methodology for sentiment analysis developed in the Netherlands on data from Mexico. We also obtained an R program for treating smart meters data from the United Kingdom and applied it to Canadian data with relative ease.

• SOS can be naturally achieved regarding sharing of methods and datasets, thanks to the approach based on the shared computation environment. However, this centralized architecture does not comply with the Common Statistical Production Architecture (CSPA) vision regarding the sharing of tools. A partial adoption of the CSPA vision could be put in place, thanks to the fact that Big Data tools are de facto standardized and implementing them on different platforms should not be an issue.

• 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 is showing for the first time, on a practical basis and on a broad scale, the potential and the limits of the use of Big Data sources for computing statistics. Improvements in efficiency and quality are possible by replacing current data sources with, for example, smart meter data or scanner data. New products can be obtained from novel sources such as traffic loops, mobile phones and social media data. However, sources can be of low quality and require some serious pre-processing before being used (e.g. web sources). In general, Big Data sources can be effectively used as additional sources, benchmarks or proxies.

• The possibility of relying on a shared environment for the production of statistics is severely limited by privacy constraints on data sets. These constraints often limit the personnel authorized to access and treat the data, and do not allow files to be moved outside the physical boundaries of a single organization.

• These limitations can be partly bypassed through the use of synthetic data sets. A synthetic data set can be obtained by perturbing a privacy-sensitive data set so that it loses any links with entities of the real world, maintaining sufficient resemblance to the real data to be considered statistically meaningful. Another solution is to generate the data by modeling its behavior. We used both approaches in the project, for smart meters data and scanner data respectively.

Findings from Teams

In the following we present a summary, specifically highlighting how the findings address the specific objectives for Sandbox stated above. Please refer to individual experiment reports for details on the activity.

CPI

Tools used: Hive, Pig, MonetDB, Spark

| SO1 Sources collection and manipulation | ❗ It was not possible to install real data due to privacy issues, so we had to rely on synthetic data. We created a simple model useful for testing performance but we needed real data to have statistically meaningful data. Data can be manipulated effectively with Big Data tools (several of them were tested). |
Scanner data are already used in production in several countries and others are currently studying how to incorporate them.

Scanner data are a superset of data currently used in the CPI survey, so they can cover the current mainstream production. Moreover, they potentially allow for the production of other statistics. Scanner data are likely to have same format in every country (expect for classifications) so statistics can be comparable and methods can be reused. Particularly good case for shared environment.

Technology: Big Data technologies are effective when applied to really big datasets in the order of hundreds of gigabytes. On small-medium dataset sizes (up to a order of some gigabyte) the inherent performance overhead of starting a MapReduce jobs makes response times higher by several orders of magnitude with respect to traditional tools. The simplest query in Hive or Pig takes at least 30 seconds even on a small dataset, where a relational database produces a response in some milliseconds. On the other hand, with big datasets the response times with Hadoop MapReduce scale very well (10 times the time when increasing the data size by a factor of 1500), while cannot be handled in memory with statistical tools and imply long loading times with a database.

Technology: Different flavours of Big Data tools can be used effectively to speed up data analysis. Pig and Hive allow to write processing scripts in a very compact form, although there is a learning curve involved for taking maximum advantage from the technology. RHadoop forces the user to adapt to the MapReduce paradigm, which is not always the most intuitive form for writing analysis scripts and can lead to long and complex programs. On the other hand it gives more control to the developer as well as the possibility of using the entire R library.

Technology: We experienced some issues with the MapReduce framework when running a particularly complex query on a big (about 250Gb) dataset. The problem was quite difficult to interpret and could be solved only by relying on the Hortonworks support. This was an example of possible issues that could arise when working with an Hadoop infrastructure in production, suggesting the need of a professional support.

Statistics: Scanner data and web scraping are already used in several statistical organizations in production. A more widespread use can involve new products and more timely and efficient computation of traditional products as well as standardized treatments. Moreover, it can help statistical organizations to respond to the “menace” of prices index computed by both research projects and IT companies.

Privacy: Scanner data are highly sensitive data and they cannot be shared in the Sandbox. However, they can be modeled fairly easily. A more refined synthetic data generator can allow sharing of techniques and results, which can be applied on real data within statistical organizations.

Partnership: The work on web scraping in this team and in the job vacancies team showed that it is difficult to gather a representative amount of data scraped from the web. While working on this team, a global scraped data provider was discovered, www.pricestats.com. An attempt to establish a partnership with this company could be made in order to acquire large amounts of scraped data more easily. Partnership on a large scale involving several statistical organizations could lead to favourable conditions and shared solutions for data processing.

Skills: Trained IT professionals could acquire knowledge on 3-4 different languages during project lifespan. This “dynamic” approach is required because the offering is constantly evolving and even the same tool can present significant differences from one version to another (e.g., Spark)

Mobile phones
Tools used: MonetDB, Pentaho, Hive

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Tokens used: MonetDB, Pentaho, Hive</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO1 Sources collection and manipulation</td>
<td>✗ Impossible to share real data at detailed level due to privacy constraints. Relying on data from Orange, difficult to collect. Data was manipulated with different tools.</td>
</tr>
<tr>
<td>SO2 Production of quality statistics</td>
<td>✗ The data is anonymized and aggregated so it is not clear at which level it can be used for statistics.</td>
</tr>
<tr>
<td>SO3 Correspondence with existing products</td>
<td>✓ Several studies already show that mobile phone data can be used as an auxiliary or primary source. Experiments in the Sandbox with Orange data are in progress.</td>
</tr>
<tr>
<td>SO4 Cross-country sharing</td>
<td>✗ Common statistics require agreement on the data level. Output from future work can be recommendations of minimum requirements for producing statistics. Comparisons with the Slovenian case are in progress.</td>
</tr>
</tbody>
</table>

Technology: Some query seem to cause problems for MonetDB: answers requires many minutes, strange errors appear, CPU and memory usage very high.

Methodology: Often Big Data need additional information; in our experiments we had data about mobile antennas but we didn't know “where” they were placed: in the city or in the country, in the capital or in which region. So we added this kind of information derived from Google Maps, obtaining much more information from our data.

Privacy: Terms of Conditions for data usage were analysed and compared with Slovenian one.

Quality: The Orange datasets can be a good starting point for working with mobile phone data, but they are too aggregated and moreover they represent a reality far from ours. This can happen in general when it is not possible to access "raw" data from the telephone company.
Partnership: Acquisition of data from providers can take a lot of time! Although there were no substantial obstacles involved, the whole process of acquisition, from the initial contact to data availability, took almost 5 months.

Traffic Loops

**Tools used:** RHadoop, R

| SO1 | Sources collection and manipulation | Difficult to move the dataset from data provider to the Sandbox, due to big size. Needed to physically ship a disk. |
| SO2 | Production of quality statistics | Quality of the source |
| SO3 | Correspondence with existing products | Data will be used to release a novel product in the Netherlands |
| SO4 | Cross-country sharing | Not clear if this type of dataset is available in all countries. If so, and data is public, share methods and environment |

Technology: RHadoop provided a framework for working in an efficient way, executing the collection of the information and posterior analysis and processing in a single step.

Technology: There is still room for improvement in the Hadoop part. Specially in what concerns the management of errors which is not well enough documented and results are sometimes very hard to interpret.

Technology: When working with RHadoop, it is therefore good to test the scripts in stand alone R, creating test scripts that generate the key-value pairs.

Technology: Since R is not the most efficient language when it comes to processing speed, it seems that implementing computationally intensive methods should not be done in RHadoop.

Statistics: Although more sophisticated indicators may be more appropriate for the particular traffic phenomenon, the simple aggregate indices could be suitable for measuring the evolution of flow variables that count the occurrences of interesting phenomena such as, for example, people walking by strategic tourist points.

Quality: The quality of the data is poor. A substantial portion of the minute data is missing and can be reconstructed using the large redundancy in the data.

Skills: Important skills that were needed for this experiment were especially data science skills. Also, a lot of IT skills for writing programs to manage the data, and a lot of expert knowledge and methodological knowledge were used to come to the right results. As described in the blog of Drew Conway ([http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram](http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram)), first the IT skills are used or formed, after which the statistical/methodological skills are used, after which one can start to think about the substantive knowledge. This was also the experience we had: first one starts to think how the data should be tamed in an IT sense, then a lot of methodological work is done, after which an analysis of the data can really start.

Smart Meters

**Tools used:** Pig, R

| SO1 | Sources collection and manipulation | Privacy constraints, needed synthetic or anonymized data. |
| SO2 | Production of quality statistics | The source can provide more accurate, relevant, coherent and timely statistics on power consumption than usual sources. Big Data tools are necessary for dealing with the size of the datasets. The fine level of granularity allows correlate with other data, e.g. weather, price. |
| SO3 | Correspondence with existing products | Power consumption and power price related survey. Possible use as auxiliary source in the Census. |
| SO4 | Cross-country sharing | The data format is similar. We could use script from the UK and apply it to Canadian data. Same script for aggregation. |

Technology: Pig was used to quickly generate aggregations on cleaned data. Irish dataset (150 million records - 2.5Gb) was aggregated at day level in 2 minutes.

Technology: Pig does not natively define statistical functions. This was easily solved using a third-party extension (developed and freely made available by LinkedIn) with statistical functions. User Defined Function can indefinitely extend the power of the Pig language and there are several useful functions freely available.

Quality: Evaluation of quality framework has been provided (see experiment reports for details)
Partnership: In Canada currently, only local distribution companies (LDCs) and their authorized agents have the ability to transmit or request information from the Meter Data Management and Repository system, which processes and stores smart meter data for a geographic location. Thus, in order to obtain smart meter data, we first need to obtain data sharing agreements with LDCs.

Skills: Based on our experiences during this project, we found that working with Big Data requires enthusiastic persons with the following skills:

- Methodological (e.g., understanding of statistical techniques, mathematical modelling, data mining and machine learning).
- Technical (e.g., programming experience in languages such as PigLatin, Hive, Java, Python, Ruby, SQL, SAS, R). But ultimately, the programming skills needed are dependent on the data and tools available and the type of problems that need to be solved.
- Creative (since there is never a "how-to" guide on the many potential challenges and opportunities that come with new Big Data sources).
- Communicative (as one would need to liaise between IT, methodology, subject matter and stakeholders).
- A passion for learning as one would need to continuously upgrade one's skills.

Social media

Tools used: Spark, Hive

<table>
<thead>
<tr>
<th>SO1 Sources collection and manipulation</th>
<th>Collection of meaningful data required a lot of time (months). Twitter public API is always limited.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO2 Production of quality statistics</td>
<td>Quality in general is an issue. for sentiment, many unused records or difficult to classify.</td>
</tr>
<tr>
<td>SO3 Correspondence with existing products</td>
<td>Tourism statistics. Sentiment correlated to consumer confidence from Dutch studies.</td>
</tr>
<tr>
<td>SO4 Cross-country sharing</td>
<td>Experiment of application of the Dutch methodology to Mexican data. Possible sharing of the method but results were not very comparable - sentiment is related to language and using just emoticons is not sufficient.</td>
</tr>
</tbody>
</table>

Statistics: The same data source could be used for two different statistics.

Statistics: Different studies on mobility could be carried out using Twitter data.

Methodology: Application of cross-national methodology (Dutch methodology applied to Mexican data).

Methodology: Could not show a correlation between tweets and Mexican consumer confidence using the Dutch methodology. Possible explanations: i) Twitter dataset may not completely contain all public Mexican tweets created during the period studied; ii) only messages containing emoticons were used to determine changes in sentiment; iii) in the Dutch study the sentiment in Facebook messages were found to be highly correlated while Twitter alone correlated much less.

Job Vacancies

<table>
<thead>
<tr>
<th>SO1 Sources collection and manipulation</th>
<th>We derived a cheap and fast way to collect job ads data from Italy with more companies than currently in the sample.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO2 Production of quality statistics</td>
<td>Quality of the sources is in general low. However, it is easy to expand sources. Corresponding statistics more timely and granular.</td>
</tr>
<tr>
<td>SO3 Correspondence with existing products</td>
<td>Job vacancies indicator requires another variable that is not possible to collect through the web but should be taken from administrative sources. We could not compare the trend because of missing data. (next year result)</td>
</tr>
<tr>
<td>SO4 Cross-country sharing</td>
<td>Cleaned data is in the same form so methods can be shared. Cleaning phase is country-specific.</td>
</tr>
</tbody>
</table>

Statistics: The number of sources was limited by the capability of the tools and the structure of the web sites

Statistics: The sources could not guarantee all the variables that are necessary for computing the indicator. The source can be used for a different, simplified, indicator or it has to be integrated with other sources, or used as a benchmark.

Quality: Not easy to assess the coverage of the source
The capability of HDFS to handle clusters of machines makes it a good storage solution when data grows constantly and scalability of Big Data tools comes at the price of complexity of installation and management. Specialized IT support is required for set up and maintenance of the infrastructure. In a production environment, an upgrade protocol should be established in order to keep up with the latest versions of the frequently changing tools. Researchers should be aware of this and be prepared for working in “unfriendly” environments, as well as considering frequent switches from one tool to another. In a production environment, an upgrade protocol should be established in order to keep up with the latest versions of the frequently changing Hadoop platforms. In general, the skills required are a mix of system managers and architect-engineers.

4. Training and dissemination

4.1. Training

Skills related to Big Data are still lacking in organizations and this is considered a factor that can limit the planning and initiation of new projects on Big Data. Two surveys focusing on the use of Big Data in statistics, respectively promoted by UNSD and UNECE, report this as a critical aspect for organizations:

<table>
<thead>
<tr>
<th>SO1 Sources collection and manipulation</th>
<th>✔ Could effectively extract the required sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>SO2 Production of quality statistics</td>
<td>❓ Still not clear if and how an approach based on massive automated extraction of web data can support official statistics</td>
</tr>
<tr>
<td>SO3 Correspondence with existing products</td>
<td>❓ Ditto</td>
</tr>
<tr>
<td>SO4 Cross-country sharing</td>
<td>❓ Language differences can make difficult the application of text mining techniques. Data extraction techniques can be easily shared, though</td>
</tr>
</tbody>
</table>

Technology: Trade off between automatization of the collection procedure and structure of the collected data (write better)

Technology: During the experiment 2000 websites were scraped at depth level 2 in 2 hours and 10 minutes using the cluster computing and storage capabilities; the same task took 8 hours and 40 minutes on a single server; so a considerable improvement of performance was experienced (73.41% reduction in time spent).

3.3. Technology

**Big Data Architecture.** One of the key questions behind the project was related to testing the use of tools for Big Data processing for statistical purposes. This is a novel approach for statistical organizations, though these tools were only used in an experimental context. Hence, some experiments were specifically targeted at testing the performance of Big Data tools over “traditional” tools such as relational databases and statistical software. The results of our experiments are that Big Data tools such as Hadoop allow us to overcome the limits of memory and computing power of single machines, when datasets grow in the order of tens and hundreds of gigabytes. “Traditional” tools, especially if well-tuned and running on servers, still provide better performance when working on dataset sizes up to single gigabytes.

This suggests that the ideal architecture for Big Data processing is a hybrid one where Big Data tools coexist with traditional ones: the Hadoop infrastructure can be used as a solution for data storage and preliminary analysis, while traditional tools can be used to analyze a selected, pre-processed subset of data. More specifically, the capability of HDFS to handle clusters of machines makes it a good storage solution when data grows constantly and indefinitely. For example, in a continuous data collection like scanner data, it is difficult to fix a higher limit of space required for storing data. HDFS can allow to dynamically expand the storage capabilities by adding nodes to the cluster. This technical capability has an important implication for data management choices: organizations can afford not to throw away data which is not immediately “useful”, keeping a warehouse of historical data that is always accessible for analysis (through a distributed programming paradigm like MapReduce or Spark) and can be downloaded at any time.

**Performance.** MapReduce (and its high-level interface, Pig, Hive and RHadoop) is in general “slow”. The simplest query requires at least 30 seconds. This is the minimum time needed by the infrastructure to setup the distributed computation (including for example the activation of the nodes and the collection of results). However, this is a fixed overhead that becomes less and less significant as the size of data grows. The key to achieve scalability is to split the datasets on different machines and let each machine perform the work on the part of data it hosts. So, the larger the dataset, the more nodes involved in the computation, the higher the throughput.

**Management.** The scalability of Big Data tools comes at the price of complexity of installation and management. Specialized IT support is required for set up and maintenance of the infrastructure. The time required for setting up the infrastructure by ICHEC resources was approximately one month. After that, a system administrator has continued followed the project in order to manage user creation and platform maintenance. We contacted the support from Hortonworks 6 times. In particular in the early phase of the project, their help has been particularly useful to fix some configuration problems. However, once the solution has been found, the acquired know-how allowed to handle similar situations without relying on the support. We also found out that several issues were solved in an updated release of the platform, but we chose not upgrade in order not to interrupt the project activity.

In general the tools used were fairly “unstable”. Some tools were added and several updates were required even in the short time frame of the project. Researchers should be aware of this and be prepared for working in “unfriendly” environments, as well as considering frequent switches from one tool to another. In a production environment, an upgrade protocol should be established in order to keep up with the latest versions of the frequently changing Hadoop platforms. In general, the skills required are a mix of system managers and architect-engineers.

**Data collection.** When large data sizes are involved, data acquisition itself can be a difficult task. The traffic loops activity team had to load in the Sandbox the full dataset, which amounts to 3Tb. This would have taken weeks to load through FTP, so a disk had to be physically shipped to ICHEC’s data center.

Web Scrapping

**SO1 Sources collection and manipulation**

Could effectively extract the required sources

**SO2 Production of quality statistics**

Still not clear if and how an approach based on massive automated extraction of web data can support official statistics

**SO3 Correspondence with existing products**

Ditto

**SO4 Cross-country sharing**

Language differences can make difficult the application of text mining techniques. Data extraction techniques can be easily shared, though
At present there is insufficient training in the skills that were identified as most important for people working with Big Data. Skills on Hadoop /NoSQL DBs indicated as “planned in the near future” by majority of organizations." (from UNSD global survey on the use of Big Data)

“Projects in planning were less likely to use tools generally associated with “Big Data”. Often this decision was made due to a lack of familiarity with new tools or a deficit of secure “Big Data” infrastructure (e.g. parallel processing no-SQL data stores such as Hadoop).” (from skills survey of the UNECE Modernisation Committee on Organisational Frameworks and Evaluation)

The Sandbox environment allowed to tackle this issue by providing a shared environment for testing and building new competencies in Big Data without direct financial costs for organizations on hardware and software licences. The collaborative approach allowed to gain a technical know-how in a relatively short time and to create a community of Big Data experts involving both statistical and IT resources. Two training sessions were organized during two face-to-face project workshops in Rome (May 2014) and in Heerlen (September 2014). The training included both a general presentation of the technical topics and a specific focus on the tools included in the Sandbox. Participants could test the tools directly on the Sandbox and share their knowledge among each other. The material produced for the training is publicly available on the wiki and will be updated and expanded in the continuation of the project in 2015. Findings on technology (reported in the previous section) also can provide practical indications for organizations that need to set up Big Data infrastructures.

The outcome of the skill-building approach tested in the project was positive: all tools available in the Sandbox were used in the experiments by both researchers and technicians having no former specific experience. The relevant skills could then be acquired starting from the training and the mutual collaboration, using the Sandbox as a common learning platform. Following this result, we believe that the Sandbox can represent in the future an effective capacity building platform for organizations.

4.2. Dissemination

Project results will be disseminated in the context of a satellite event to the NTTS 2015 conference in March 2015. This will feature general presentations of the project, presentations of the results from experiment teams and discussions on future work. Furthermore, 5 abstracts focusing on work made in the context of the project have been accepted for presentation in the main track of the NTTS conference.

Project deliverables will be available on UNECE Wiki. List of final deliverables:

- Big Data Project Summary Report (this document)
- Partnerships Task Team
  - Partnerships abstract for NTTS March 2015.docx
  - Guidelines for the establishment and use of partnerships in Big Data Projects for Official Statistics
- Privacy Task Team
  - NTTS abstract on results Privacy Task Team.docx
  - Deliverable 1: Stock taking exercise
  - Deliverable 2: Big Data characteristics and their implications for data privacy
  - Deliverable 3: Big Data Privacy
- Quality Task Team
  - Big Data Quality Framework Abstract for NTTS.docx
  - Big Data quality frameworks UNECE task team outcomes v1.5.docx
  - Big Data Quality Framework - final.pdf
- Sandbox Task Team
  - Sandbox Team NTTS Abstract
  - Experiment report: CPI Team
  - Experiment report: Canadian Smart Meter Data
  - Experiment report: Job Vacancies
  - Experiment report: Mobile Phones
  - Experiment report: Social Data
  - Experiment report: Social Data: Mobility Studies
  - Experiment report: Traffic Loop data
  - Experiment report: Web Scraping

5. Project Management and Resourcing

5.1. Project Execution

Big Data project was managed in line with the HLG agile project management approach. The UNECE project manager and two coordinators were responsible for coordinating all project activities and regular reporting to the Executive Board. The project sponsor, the HLG, has ultimate responsibility for signing off the project deliverables. In practice, this responsibility is delegated to the Executive Board.

Timeline of key milestones in project execution:

- Big Data virtual sprint (March 2014): 16 participants from 12 national and international statistical organizations focused on finding agreement on the major strategic questions posed by the emergence of Big Data and developed a discussion paper “How big is Big Data? Exploring the role of Big Data in Official Statistics”.
- Big Data Project Workshop Rome, Italy (2-3 April 2014): 17 participants from 13 national and international statistical organizations identified six main project priorities and recommended the establishment of project teams for four (Quality, Privacy, Partnerships, Sandbox) and Modernisation Committees for remaining two priorities (Skills, Inventory).
- Workshop on the use of Big Data for official statistics held in Dublin Castle, Ireland (16 April 2014): approximately 40 participants attended a presentation on the project, including the official launch of the “Sandbox” environment.
• Establishment of task teams (May 2014): detailed project structure with project manager, two coordinators and four task teams; altogether approximately 75 individuals.
• Big Data Sandbox project training in Rome, Italy (28-30 May 2014): 20 participants from 12 national and international statistical organizations attended training on Big Data technologies and practically tested new tools in Sandbox environment. Seven experiment activity teams and their scope of work were established and started their work.
• Big Data project workshop in Heerlen, Netherlands (9-12 September 2014): 16 participants from 11 national and international statistical organizations presented preliminary findings from all task teams and agreed next steps including dependencies between task teams and project reports. One day was dedicated to additional training.
• Big Data Questionnaire developed in partnership with the United Nations Statistical Division (September - October 2014): collected information from 33 organizations, and about 57 Big Data projects.
• Virtual Sprints of task teams Quality (September 2014) and Partnerships (October 2014): finalising deliverables including analysis of results of Big Data Questionnaire.
• Development of abstracts for NTTS (October 2014): all four project teams prepared two page summaries of their work and submitted paper abstracts to the scientific committee for the New Techniques and Technologies for Statistics (NTTS) Conference 2015. All submissions have been accepted and there was also agreement to organize a parallel HLG Big Data event.
• Workshop on the modernisation of statistical production and services, Geneva, Switzerland (19-20 November 2014): task teams presented their work, findings and recommendations to the workshop participants. Work of task teams received strong support from participants resulting in proposal to continue work on Big Data in 2015.
• Finalization of all deliverables (December 2014): task teams finalised all deliverables.

5.2. Project Resources

The following resources were used:

1. Start-up phase:
   a. Virtual sprint: 4 virtual sprint sessions with 10 participants, 3 hours including preparation: 120 person hours
   b. Sprint in Rome: with 17 participants for 2 days, 6 hours on average per day plus 5 hours for preparation: 289 person hours
2. Task Teams:
   a. Privacy: 14 meetings with 7 participants on average, 2 hours including preparation: 196 person hours
   b. Partnerships: 13 meetings with 5 participants on average, 2 hours including preparation: 130 person hours
   c. Quality:
      i. 14 meetings with 9 participants on average, 2 hours including preparation: 252 person hours
      ii. 6 virtual sprint sessions with 6 participants, 4 hours including preparation: 144 person hours
   d. Sandbox:
      i. 15 meetings of full Sandbox team with 12 participants on average, 2 hours including preparation: 360 person hours
      ii. 8 meetings of Sandbox Core team with 8 participants on average, 2 hours including preparation: 128 person hours
      iii. 2 face to face workshop sessions with 23-16 participants resp, 6 hours on average per 3-4 days per ws resp, plus 5 hours preparation per ws: 993 person hours
      iv. 7 Experiment teams with 5 participants on average, 3 hours per week for 24 weeks of practical work in Sandbox environment: 2520 person hours
3. Coordination:
   a. Coordination meetings and minutes for Task Teams: 200 person hours
   b. Summary report and NTTS abstracts: 150 person hours
4. Contribution to HLG workshop in Geneva, November 2014: 4 participants with 20 hours including preparation: 100 person hours

Total: 5582 person hours (just over 3 person years)

5.3. Project Reporting

The project generally progressed according to the plan without any critical issues. Achievements, risks and issues have been reported to the Executive Board on a monthly basis.

References


