ONS Big Data Project – Progress report: Qtr 3 July to Sept 2014

Jane Naylor, Nigel Swier, Susan Williams, Karen Gask *Office for National Statistics*

Background

The amount of data that is generally available is growing exponentially and the speed at which it is made available is faster than ever. The variety of data that is available for analysis has increased and is available in many formats including audio, video, from computer logs, purchase transactions, sensors, social networking sites as well as traditional modes. These changes have led to the big data phenomena – large, often unstructured datasets that are available potentially in real time.

Like many other National Statistics Institutes (NSIs) the Office for National Statistics (ONS) recognises the importance of understanding the impact that big data may have on our statistical processes and outputs. So ONS established a 15 month Big Data Project to investigate the potential benefits alongside the challenges of using big data and associated technologies within official statistics. This is due to complete at the end of March 2015. The key deliverable from this proof-of-concept project will be an ONS strategy for big data. In taking forward this work ONS is upholding all relevant legal and ethical obligations.

Summary

This report provides an overview of progress on the ONS Big Data Project during the third quarter (July – Sept 2014) and builds on the work that was documented in the first and second quarter progress reports[1]. An update is provided on the practical elements of the Big Data project: the four pilot projects covering economic and social themes. Each pilot uses a key big data source, namely Internet price data, Twitter messaging, smart meter data and mobile phone positioning data. Their objectives will collectively help ONS to understand the issues around accessing and handling big data as well as some of their potential applications within official statistics. Alongside the pilot projects a significant activity within the Big Data Project will be stakeholder engagement and communication.

[1] http://www.ons.gov.uk/ons/guide-method/development-programmes/the-ons-big-data-project/index.html

Contents

1 Introduction

The high level aims of the ONS Big Data Project are to:

- investigate the potential advantages that big data provides for official statistics; to understand the challenges with using these sources; and to establish an ONS policy on big data and a longer term strategy incorporating ONS's position within Government and internationally in this field; and
- make recommendations on the best way to support the ONS strategy on big data beyond the life of this project.

A major component of the project is to include some practical applications of big data, to both assess the role they might have within official statistics and to help understand the methodological, technical and privacy issues that may arise when handling them.

Four pilot projects have been chosen, covering economic and social themes. Each pilot uses a different big data source, namely Internet price data, Twitter messaging, smart meter data and mobile phone positioning data.

Although ONS is researching only samples of these data, even these can be too large and complex to process efficiently using standard ONS computers. The solution is to use the ONS innovation labs, a private 'cloud' based environment, for analysing them.

This report briefly introduces the ONS innovation labs, then provides an overview of progress on the four pilot projects in the second quarter (July – Sept 2014). In addition a summary of progress around stakeholder engagement for the project is provided, an important activity for the project. This report builds on the work that was documented in the first and second quarter progress reports[2].

In all these activities ONS is committed to protecting the confidentiality of all the information it holds. In order to produce statistics using big data sources we are interested only in trends or patterns that can be observed not in data about individuals. However, we recognise that accessing data from the private sector or from the internet may raise concerns around security and privacy. The Big Data Project is therefore accessing only publically available, anonymous or aggregated data and this data will be used only for statistical research purposes. In addition all of our work fully complies with legal requirements and our obligations under the Code of Practice for Official Statistics. During this quarter we have also developed an ONS Big Data Policy that sets out operational principles and guidance for using big data within government, which aims to address these concerns.

2 Innovation labs

The ONS innovation labs have been set up to help facilitate research into new technologies and open source tools, new sources of public data, and to develop associated skills. The innovation

[2] http://www.ons.gov.uk/ons/guide-method/development-programmes/the-ons-big-data-project/index.html

labs are a key enabler for the ONS Big Data project because they allow us to handle large and complex data sets and to test new big data technologies.

These labs consist of a number of high-specification desktop computers with some additional network storage. The hardware is configured using OpenStack[3] technology. This provides a very flexible environment to deploy different 'virtual environments' depending on the processing and storage requirements of different projects. In particular, this approach will provide a flexible framework for experimenting with big data parallel computing technologies such as Hadoop[4]. The innovation labs have been designed to provide a route for accessing open source tools.

We have placed restrictions on the data that can be accessed in the labs. In the Big Data Project these are currently confined to the Twitter and internet price data pilots, which are using publicly available data, and the analysis of anonymous smart meter information.

The innovation labs have been utilised extensively to focus on data analysis. As this has progressed, and the data sizes increased, additional methods for processing large datasets have been explored. The team have started to explore real data science - smart ways for processing data through understanding the IT environment. Implementation of a fully working Hadoop cluster has not yet been possible. A single node Hadoop instance has been used for training and exploratory purposes, but this needs to be expanded so that the performance gains from running Hadoop across multiple machines can be tested.

3 Prices pilot

Background

Web scrapers are software tools for extracting data from web pages. The growth of on-line retailing over recent years means that many goods and services and associated price information can be found on-line. The Consumer Price Index (CPI) and the Retail Price Index (RPI) are key economic indicators produced by ONS. Web scraping could provide an opportunity for ONS to collect prices for some goods and services automatically rather than physically visiting stores. This offers a range of potential benefits including reduced collection costs, increased coverage (ie more basket items and/or products), and increased frequency.

Supermarket grocery prices have been identified as an initial area for investigation because food and beverages are an important component of the CPI and RPI basket of goods and services.

Research objectives

The objectives are to:

- Set up and maintain prototype web scrapers to test the technical feasibility of collecting price data from supermarket websites.
- Develop methods for quality assuring scraped data.
- compare scraped data with data collected using current methods, explore methodological issues with scraping prices from supermarket websites

[3] http://www.openstack.org/
[4] http://hadoop.apache.org/

- Establish whether price data could be sourced directly from commercial companies and if so, how these compare with data scraped by ONS prototypes.
- Evaluate the costs and benefits of these alternative approaches to collecting price data

Progress

Data collection

The daily web-scraping operation for a selection of 40 item categories and three on-line supermarkets continued during the quarter with some further development and refinement of the quality assurance processes to improve anomaly detection.

The experience of this pilot so far suggests that a web-scraping operation of this scale is a cost-effective means of collecting large volumes of data. As long as appropriately skilled resource is on hand to fix occasional problems then data can be collected with minimal overheads. However, it is important to stress that this pilot has involved scraping only three websites; maintenance overheads would increase with an operation scraping a larger number of websites.

For several weeks the scrapers were run in parallel across two different locations with the aim of testing whether they collected the same data. This test did not identify any differences in price data, although there were some small differences in item availability. This suggests that these supermarkets are operating national pricing policies for on-line groceries, although different locations may hold different items in stock.

Data analysis and methodology development

Good progress has been made with the initial analyses and evaluation of these data. The pilot has been working with the University of Huddersfield who have completed an internal assessment and high level review of the methodological implications of using these data for price statistics. These early findings point to a number of ways in which the much larger volume of data available through scraping methods could deliver various improvements. However, the sheer volume of data also raises methodological questions about the best way of utilizing it as there are numerous options.

Commercial data purchase:

The pilot team have had useful discussions with MySupermarket.com and an agreement has been reached on purchasing 3 years of daily price quote data for a selection of item categories and supermarkets. To emphasise this is one way transaction of data, where appropriate ONS will be purchasing data for analysis within the Big Data Project but we would never consider selling on our findings or data.

One of the main challenges in creating price indices from bulk web scraped data is the linking of products across time. This is important because price changes of the same or equivalent products are a fundamental concept on which price indices are constructed. For example, if a price quote has been collected for a particular brand of tinned tomatoes in one month, we will get a more accurate measure of price change if we are able to collect the quote for the same brand of tinned tomatoes from the same store than if we collect data for some other brand.

This linking of products is made even more challenging by the high rate of product turnover of supermarket products, typically around 30 per cent a year. These turnover patterns are often complex with product lines being discontinued and then sometimes later reintroduced. Another issue is with cosmetic repackaging, which normally involves the issuing of new product codes. While scraping price data is relatively straight-forward, linking thousands of products in a highly dynamic environment across time is much more challenging.

MySupermarket.com is a price comparison website that scrapes on-line supermarket websites and then matches the same or equivalent products so users of the website can identify the cheapest place to buy a product, or 'trolley' of products. It is the ability to match equivalent data, rather than simply collect the raw scraped data, which may offer considerable potential. If it can be proved that this matching allows products to be tracked across time, than purchasing these data might be a better long-term option, then scraping the data ourselves. Agreement has been reached on purchasing three years of daily price quote data for a selection of item categories and supermarkets.

Future work

The main focus over the next quarter will be preparing a report incorporating all aspects of this research, planned for publication in early 2015. This report will focus on bringing together all the pilot research findings to date and will not aim to cover MySupermarket.com data in any detail. Instead, this will form a major part of the next phase of research.

4 Twitter pilot

Background

Twitter is a micro-blogging site which has become one of the leading social networking platforms. Most tweets are public data and Twitter provides open source tools for accessing these data (albeit with some limits). Twitter provides an option for users to identify their current location. This means that tweets from a subset of users can be tied to specific locations over time. This data can then be used to track mobility patterns (eg Halwelka et al 2013).

A historic weakness of England and Wales mid-year population estimates has been capturing the internal migration of students. Students typically move to different parts of the country when they commence studies and then move to a new location again when they graduate and find employment. The main source for estimating internal migration is the GP patient register but young people, especially young men, are often slow to re-register when they move .

Because these populations are more likely to use Twitter than other populations (Koetsier 2013). the primary aim of this research is to determine whether geo-located data from Twitter can provide fresh insights into internal migration within England and Wales and whether these insights could be used to improve current estimation methods.

Even though these data are all publicly available, the pilot team is very conscious of the ethical issues around how these data will be used and will handle the data appropriately. Although we are working with data at the individual level (which is publicly available) our research question and ultimate interest is around patterns and trends in internal migration at the aggregate level, eg for groups within the population such as students in a particular city.

Research objectives

The objectives are to:

- Develop an application to harvest geo-located tweets from the live Twitter stream.
- Develop a method for processing these data by user to identify clusters and to derive different cluster types (ie home, work, study, and commutes).
- Develop a method for detecting changes in cluster patterns over time that could be interpreted as internal migration.
- Compare these results with current internal migration estimates and census data to understand their coverage and any resulting bias, and to establish whether these data are useful.
- Identify any big data technologies that may be needed if this research is to be taken forward over the longer term.

Progress

Data collection

The data collection method for this pilot was changed during the last quarter. We previously ran a Twitter harvesting application that collected around 500,000 tweets a day. To ensure compliance with Twitter's developer 'Rules of the Road'[5] (we were pushing the boundaries with volume and also geo-located data) the decision was made to halt the application on 15 August and purchase future data through GNIP, a company owned by Twitter offering social media data for purchase.

We can use the data collected between 10 April and 15 August but we will supplement this with data purchased from GNIP from between 16 August to 31 October. This should provide sufficient data to explore the feasibility of establishing residence and moves of the geo-located Tweets.

Data processing

One of the main challenges of this pilot is how to organise and make sense of the large amount of data collected. Of greatest interest are clusters of tweets by user. These are likely to indicate locations of some significance, such as a person's home, their place or work or study, or some other place where they spend a lot of time. In terms of measuring internal migration, these locations are important because they are most likely to indicate where someone is usually resident. Of lesser interest are locations with one-off or infrequent activity. These can be considered 'noise' points that should be filtered out of this analysis.

The clustering method adopted is a variant of DBSCAN, or 'density-based spatial clustering of applications with noise' (Backlund et al, 2011). The pilot has built on the progress of the second quarter by scaling up this algorithm to process a larger number of records. This has proved

[5] https://dev.twitter.com/overview/terms/rules-of-the-road Accessed 2 Oct 2014

challenging because of an exponential processing issue, which is a well documented problem with DBSCAN (Tsai & Wu, 2009). A method has been developed that addresses this problem through pre-processing of the data, which resulted in reasonable processing times.

The pilot has successfully developed a method of using AddressBase[6] to classify clusters by type (eg residential or commercial). The main purpose of this processing step is to help distinguish residential addresses from other types of address; this would then enable subsequent analysis to give precedence to residential addresses.

Future work

The remainder of this project will continue via two main phases.

Phase 1 will focus on analysis of data collected between 10 April and 15 August 2014. This should be sufficient to determine the feasibility of using these data to identify place of usual residence and to undertake some initial analyses, such as calculating penetration rates by local authority and then comparing these to the demographic profile of local authorities.

Phase 2 will follow, focusing on procuring additional data (from 15 August to 31 October), processing this data and then undertaking a more complete analysis, including identification of potential moves and the feasibility of producing internal migration estimates from these data.

The pilot team has had some initial discussions with the University of Southampton Social Sciences Department on providing additional analytical support for the pilot in the form of methodological support and advice.

References

Backlund H, A. Hedblom, N. Neijman, 2011, Linkopings Universitet, “DBSCAN - A Density-Based Spatial Clustering of Application with Noise” Available at: http://staffwww.itn.liu.se/~aidvi/courses/06/dm/Seminars2011/DBSCAN(4).pdf Accessed on 25-03-2014

Koetsier, J. 2013, “Only 16% of U.S. adults use Twitter, but they are young, smart and rich”. Available at: http://venturebeat.com/2013/11/04/only-16-of-u-s-adults-use-twitter-but-theyre-smart-young-and-rich/ Accessed on 18-03-2014

Hawelka, B, I Sitko, Euro Beinat, S Sobolevsky, P Kazakopoulos and C Ratti, 2013 “Geo-located Twitter as the proxy for global mobility patterns” http://arxiv.org/abs/1311.0680 Accessed on 19-03-2014

Tsai C, C. Wu, 2009, “GF-DBSCAN A New Efficient and Effective Data Clustering Technique for Large Databases”, Proceedings of the 9th WSEAS International Conference on Multimedia Systems and Signal processing”, Aavailable at: http://www.wseas.us/e-library/conferences/2009/hangzhou/MUSP/MUSP38.pdf Accessed on 15-10-2014

[6] AddressBase is the definitive source of address information within Great Britain and is available to public sector organisations under the Public Sector Mapping Agreement.

5 Smart meter pilot

Background

A smart meter is an electronic device that records and stores consumption information of either electric, gas or water at frequent intervals. These data can be transmitted wirelessly to a central system for monitoring and billing purposes.

The European Commission's Energy Efficiency Directive (EED 2012)[7] is a common framework of measures for the promotion of energy efficiency within the EU. It supports the EU's 2020 headline target on 20 per cent energy efficiency, and its provision[8] for the roll-out of smart meters requires member states to ensure that at least 80 per cent of consumers have such intelligent electricity metering systems by 2020.

The Department of Energy and Climate Change (DECC) has one of the most ambitious roll-out policies within the EU: to put electricity and gas smart meters in every home in England by 2020[9] with roll out starting in 2015.

For electricity, readings will have a minimum specification of 30 minute intervals and will be transmitted at predefined intervals to a body called the Data and Communications Company (DCC). Data access will be permitted for certain specific functions as described in legislation[10].

Smart meter electricity energy usage data is of interest to statistical organisations as it allows investigation at low levels of geography and high levels of timeliness. Additionally, within England, this data would represent an almost complete coverage of homes.

The applications of most interest for the production of official statistics are:

1. Occupancy status of homes: low and constant electricity use over a period might indicate that a home is unoccupied, which could help survey fieldwork planning.

2. Household size or structure: it is hypothesised that profiles of energy use during the day might vary by household size or the composition of a household's inhabitants.

The ultimate aim for this research is to develop methods to produce small area estimates for use within either statistical outputs or operational processes such as fieldwork. However, as a first step, it is necessary to work at an individual (yet anonymous) level to understand patterns of energy usage. Initial research proposals have been discussed with the GDS Privacy and Consumer Advisory Group and the ONS Beyond 2011 Privacy Advisory Group. If the research is successful and suggests there is real value to be had in developing these small area estimates, the privacy and ethical issues surrounding the use of these data will need increased consideration.

[7] http://ec.europa.eu/energy/efficiency/eed/eed_en.htm
[8] This provision relates to another EU Directive on smartmeter rollout (2009) which required a full cost/benefit analysis be performed prior to commencing roll-out
[9] Wales and Northern Ireland have similar policies.
[10] Legislation still being devised

Research objectives

The objectives are to:

- Understand the big data technical/methodological challenges of handling this type of data
- Assess some of the quality aspects of smart meter type data and to form ideas on how to approach further analysis. For example, how to deal with missing values etc.
- Produce higher analysis: to focus on smart meter profiles for determining occupancy status. Less priority to be given to household size/structure or data-led analysis such as a cluster analysis (dependent on data handling restrictions and analyst resource availability)
- Review research studies in academia and other NSIs.
- Research the ethical and public perception issues surrounding this type of data
- Identify the cost/benefit to ONS for using smartmeter data in specific applications
- Propose future use and further ONS research with this type of data (final report)

Progress

Data collected during consumer trials of smart meters have previously been sourced from the Irish Social Science Data Archive and loaded into the innovation labs. Around 4000 residential homes are included in these data, and anonymised samples of these were taken to start preliminary analysis to understand the data and to help identify methods of analysing them.

The focus of the research is to assess if these data can identify unoccupied households. The rationale being that a retrospective look at smart meter data may highlight, for example, the number of homes unoccupied on census day, which might then be used to validate census results. A logical extension of this research is to identify longer term vacant properties, of great use within the production of estimates, and valuable intelligence in a census or survey operation[11].

Methods for identifying unoccupied days have been investigated to ascertain the pros and cons of each method. The variety of different energy patterns observed within the research implies that it is challenging to develop a robust method of automatically identifying unoccupied homes for a given day. The two methods which show most promise are:

1. Where the variance of the energy usage over any 24 hour period is examined.

   The home is unoccupied on a given day if:

   Variance of energy usage across the 24 hours midnight to midnight < 0.01

2. Where the average night time (1am - 5am) energy usage over the previous week is compared against the average day time (5am - midnight) usage for a given day.

   The home is unoccupied on a given day if:

$$1.1 > \frac{Mean\ day\ time\ consumption\ for\ current\ day}{Mean\ night\ time\ consumption\ for\ previous\ 7\ nights}$$

[11] ONS did not think it appropriate to identify longer term unoccupied homes as a first analysis, as the smart meter data from trials would be highly unlikely to contain such homes.

During this quarter methods for processing the full smart meter data set have been developed in R within the innovation lab; previously we were able only to analyse a small sample. Ultimately, consideration will be given to the hypothetical question of how to process smart meter data representing full England and Wales coverage of over 20 million households. This will require processing across multiple computers and the use of Hadoop to enable parallel processing.

For this analysis it is necessary to work at the individual (anonymous) household level to understand the patterns and develop generic algorithms, but the ultimate aim is to develop methods to produce estimates at an aggregate level.

Future work

Over the next three months the current state of smart meter research will be written up into a pilot report with further research ongoing, possibly using machine learning, to identify households unoccupied for a whole day. This analysis can then be extended to identify long-term unoccupied households which may be methodologically easier, although also more difficult to find examples within the trial data available for research.

A request has been made to DECC for a sample of counts of standard meters by annual usage of electricity, because it is hypothesised that low annual usage might indicate vacant properties, second homes or maybe holiday homes. Analysis will be undertaken to compare the counts against census data.

6 Mobile phone pilot

Background

Location data generated through mobile phone usage is of key interest to statistical organisations because it has the potential to inform various important aspects of population behaviour. Current research around the world is focussed on:

- Population densities – at specific times of the day and/or small geographies
- Population flows – for example the number of people who travel from area A to area B
- Tourism statistics[12] – a Eurostat funded feasibility study on the use of mobile positioning data for tourism statistics has generated research within a number of NSIs, most notably Statistics Estonia, Statistics Finland and CSO Ireland.

There are a number of features, specific to these data, that have supported this growing interest including:

- The high coverage of the population who have mobile phones (94 per cent of UK adults[13])

[12] http://www.congress.is/11thtourismstatisticsforum/papers/Rein_Ahas.pdf

- There are relatively few service providers, so any one provider might have sufficient coverage to produce reasonably representative insights of total population behaviour, reducing the effort required in approaching multiple companies.

- The growth of big data technologies and methods is allowing the service providers to do more and more with their customers' data. Since 2012 the UK's main providers - Telefonica, Everything Everywhere and Vodafone - have all embarked on initiatives to use their customers' data within the development of new data products for sale.

Historically there are many academic research projects demonstrating a use of 'call event' data, which contains location information when a customer receives or sends a text/phonecall. Of more interest is the use of 'roaming' data which is passively generated from mobile phones when they are switched on and either move between masts or send out a location reading at intervals. It is speculated that roaming data might be used to produce travel patterns from an origin to a destination location. ONS has an interest in whether this might be extended to travel patterns for 'workers' as typically produced in a census.

Research objectives

Objectives are to:

- Source aggregate data from a main UK mobile phone provider on travel patterns of workers. The emphasis here is on understanding the issues involved throughout the stakeholder engagement, negotiation and procurement stages of this 'partnership' opportunity.
- Agree a method with the service provider and monitor the issues around the collaboration.
- Compare the aggregated mobile phone data with 2011 Census data on travel to work flows to assess some of the quality aspects of mobile positioning data, and to form ideas on how to approach further analysis.
- Review research studies in academia and other NSIs.
- Research the ethical and public perception issues surrounding this type of data
- Propose future use in ONS for this type of data (final report).

Progress

Because of the ethical and privacy concerns around Government departments accessing this data ONS presented the mobile phone pilot proposal to the GDS Privacy and Consumer Advisory Group and the ONS Beyond 2011 Privacy Advisory Group. Both groups, although wary of the acquisition of individual-level data, were supportive of the use of aggregated data, especially as it is to be aggregated within the mobile phone company.

ONS held meetings with three of the main mobile phone network providers in the UK: Telefonica, Everything Everywhere and Vodafone. All these companies are actively pursuing the development of data products using geolocational information from their customers, and provided intelligence around the use of mobile phone data for statistical purposes:

[13] Ofcom facts and figures communication report 2013

- Development of the data requirement for origin-destination flows is non-trivial because of the wide range of working patterns that exist. Home location is modelled as the area where a mobile phone tends to be found at night, while work location is modelled as the location where mobile phones tend to be found mostly during the day (Mondays to Fridays). Workers who have more flexible arrangements, such as part-time, nighttime and shifts, working at multiple locations, etc, may not be easily identified with such a broad approach.

- The nearest cell tower is the basis for detecting the location of a mobile device. Because cell towers have a reach of around 300-500m in urban areas, problems arise in the detection of homeworkers or workers who do not commute a great distance. This is true for rural areas especially where cell towers may have a reach of 5 km or more.

- Key demographics such as age and sex are predominately sourced from information on contract customers, who typically represent around 50 per cent of all customers. The distribution of such contracted customers is used as a proxy for pay-as-you-go customers.

- Mode of transport, another key output from census travel to work data, is more complex to model because it involves analysis of a time series of movements for each mobile phone. This analysis tends to be outsourced to companies providing transport analytics services, although the relationships need more investigation.

Discussions are being held with Government departments, for example Department for Transport (DfT), around their use of mobile phone data to ensure a joined up approach towards any future procurement exercise.

Over the past quarter some additional research using Oystercard data has been undertaken as a precursor to obtaining mobile phone data for investigating origin destination statistics. The Oystercard system provides information on travel across the tube and bus network as well as some mainline train stations within London.

Oystercard data on the counts of journeys from an origin tube station (where an Oystercard first enters the network) to a destination tube station (where the same Oystercard leaves the network) is publicly available. Furthermore, the flows are broken down by time period including journeys made between the peak travel time of 7am and 10am. ONS used these data to see if the flows of journeys conducted in peak travel time compared well with 2011 Census estimates of travel to work for those travelling mainly by underground metro, light rail or tram.

Future work

Engagement with DfT and other Government departments will be important in the next quarter to understand their activity in the application of mobile phone data. In parallel to this continuing engagement, a specification to acquire aggregated mobile phone data will be drafted. Future steps on the acquisition of mobile phone data will depend on information obtained through the stakeholder engagement.

The Oystercard research will be finalised together with a comparison between 2001 and 2011 travel to work flows, to detect areas that have changed extensively over the 10 year period. This may inform the choice of appropriate areas to request mobile phone data.

7 Stakeholder engagement

A significant big data project activity is stakeholder engagement and communication. Stakeholder engagement activities seek to achieve the following through communication and other means:

- Engage with data users/the public to understand their concerns around the use of big data within official statistics, and their requirements for new types of outputs
- Engage with external stakeholders to acquire their data/tools/technologies for use in pilot projects
- Engage with external stakeholders to learn from their experience, to develop our knowledge and skills, co-ordinate efforts, to develop partnerships and work collaboratively with them
- Engage with internal stakeholders to co-ordinate efforts, to ensure the project's objectives align with ONS strategic objectives, and to ensure support for the project across the ONS
- Manage stakeholder expectations at various stages of the programme.

The following nine groups of stakeholders have been identified for the project:

- Privacy groups
- International
- Academia
- Private sector
- 'Big Data' companies
- Technology providers
- Government
- ONS
- Data users including the public.

In this third quarter of the project Government has been a key stakeholder identifying and contributing to collaborative opportunities and starting to develop proposals for future work. In addition engagement has increased with academics and big data companies (to raise awareness of the project, identify common interests and collaborative opportunities and learn from external expertise) and the private sector (to acquire data). Engagement has continued with international statistical organisations and privacy groups (particularly to inform the development of an ONS policy for big data).

In the fourth quarter of the project these activities will continue, with particular focus on discussing and scoping future big data initiatives across government.

Key activities in these stakeholder groups are provided below:

- The ONS Big Data team continues to engage with stakeholders across Government. The team are contributing to the Cabinet Office Community of Interest meetings to support their Data Science Programme. ONS is also contributing to a cross-profession working group focused on the capability strand of the Data Science Programme, trying to understand the gap between traditional analysis and data science, and the role the existing analytical professions should play in building capability. In addition presentations on the ONS Big Data Project were made at the Government Heads of Analysis Conference and Government Statistical Service (GSS) Methodology Symposium in June. A presentation was given at the GSS Heads of Profession meeting in September. Discussions with colleagues across Government have started to form a proposition around closer working and collaboration across Government departments.

- In addition a number of bilateral meetings/conversations/presentations have been held with representatives from different government departments in order to move forward the work of the project, share experiences and investigate collaborative opportunities:

    - engagement with Government Digital Service (GDS) and Cabinet Office colleagues around acquiring mobile phone data; internal authorisation has now been granted to proceed with a procurement exercise
    - conversations/meetings have been held with Department of Transport, to discuss use of mobile phone data
    - discussions with statisticians from Department of Energy and Climate Change around the acquisition of data to support the smart meter pilot
    - presentation/discussions held with Defence Science and Technology Laboratory DSTL, Bank of England and Department of Business Innovation and Skills with particular focus on innovation lab environment

- The main challenge around the use of big data within Government is to maximise benefits to the public while protecting the privacy of individuals. Whereas Government has traditionally collected its own data through administrative systems and surveys, most big data are produced by commercial organisations. This ranges from sources of public data available from websites through to outputs based on personal data available for purchase. Although there is potential to use these data to inform government decision making, there are new and important legal and ethical issues that need to be considered. The ONS Big Data team has developed a draft policy that pulls together a set of operating principles designed to deal with these new issues together with long-standing principles. This draft is a comprehensive policy with supporting guidance on the appropriate use of big data. A draft version of the Big Data Policy was distributed to the GDS Privacy and Consumer Advisory

Group and the ONS Beyond 2011 Privacy Advisory Group whose many useful comments were taken into future drafts.

- The Economic and Social Research Council (ESRC) have a significant amount of funding to invest in a Big Data Network to help optimise data that is available for research. Aligned to this initiative, discussions have been held with the ESRC to explore the possibility of jointly funding research into public attitudes of the use of big data for research/official statistics.

- There is a need to engage more widely with academic institutions to understand:

  - what research activities are being undertaken
  - what expertise exists
  - the skills required to work within data science
  - the types of skills new graduates have in this field, and how to recruit/attract graduates to the ONS

  An initial review was undertaken to identify UK universities offering courses on big data/data science/data analytics with follow up engagement with a subset of these institutions covering their specific courses and potential opportunities for collaboration, in particular around placement students and areas of research. More focused engagement with a smaller number of universities will be taken forward in the next quarter.

- The ONS Big Data team is also engaging with universities having specific expertise that is relevant to our pilot projects:

  - Southampton University has been commissioned to research smart meter data
  - a workshop was held with academics from the University of Cardiff with specific expertise on the analysis of data from Twitter
  - the ONS Big Data team is also engaging with academics from the Statistical Sciences Research Institute at Southampton University to provide analytical support for the Twitter pilot
  - academics from the University of Huddersfield have completed an internal assessment and high-level review of the methodological implications of using web scraped data for price statistics

- This quarter has seen significant engagement with the private sector to acquire data for use within the pilots:

  - agreement has been reached on purchasing three years of daily price quote data from MySupermarket for a selection of item categories and supermarkets
  - there has been extensive engagement with Twitter to ensure compliance with their developer ‘Rules of the Road’. Future data will be purchased through GNIP, a company owned by Twitter offering social media data for purchase
  - meetings have been held with three main mobile phone network providers in the UK: Telefonica, Everything Everywhere and Vodafone. All companies are actively pursuing the development of data products using geolocational information from

their customers, and are interested in collaborating with ONS to produce anonymised origin-destination flows of workers for comparison with census data.

- Tthe ONS Big Data Project wants to understand the big data expertise that exists within the commercial sector and how it might help achieve the project's objectives. Initial meetings have been held with a couple of big data companies to start to explore these issues:
  - Google: discussions focused on early ONS work with Google trends; contacts made for further discussions
  - Spotify: ONS had wide ranging and useful discussions around all aspects of the project; there is much potential for future engagement with Spotify.

- The main international stakeholder engagement has been participation in a 12 month UNECE international collaboration project focused on big data[14]. Members of the ONS Big Data Project are contributing to two of the task teams focused on partnerships and technology (selected because of the overlap of issues that need to be addressed in the UK context as well as internationally).

- A European Statistical System (ESS) taskforce on big data and official statistics has also been established. The taskforce is focused on the Scheveningen Memorandum[15] and its implementation through an action plan and roadmap. Members of the ONS Big Data team have contributed to the development of this action plan and roadmap, which was signed off by the European Statistical System Committee meeting in September.

8 Conclusions

This report has provided an overview of progress on the ONS Big Data Project during the third quarter of 2014. Updates on the practical elements of the Big Data project, including the ONS Innovation Labs have been provided. Each pilot project uses a different big data source and has a different set of objectives which, collectively, will help ONS to understand the issues around accessing and handling big data as well as some of their potential applications for official statistics. Alongside the pilot projects a significant Big Data Project activity is stakeholder engagement and communication. This report has also summarised key engagement and communication activities.

[14] http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics
[15] http://www.cros-portal.eu/news/scheveningen-memorandum-big-data-and-official-statistics-adopted-essc