United Nations ECE/CES/2016/16

Economic and Social Council

Distr.: General
17 February 2016

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-fourth plenary session
Paris, 27-29 April 2016
Item 3 of the provisional agenda
Strategic partnerships

Partnerships in data production

Note by the Office for National Statistics of the United Kingdom and Statistics Austria

Summary

The challenges facing the global society, economy and environment cannot be sufficiently well understood without excellent data, and these data cannot be produced without the cooperation of several actors who bring their different attributes to the task. Producers of official statistics are one of those actors and have a key role to play. Yet we can best contribute in partnership with others who have strengths that complement our own. This paper proposes that the formation of partnerships with the information industry should become business as usual for statistical offices to fulfil their core remit: better statistics, better decisions and better lives.

The paper is presented for discussion to the Conference of European Statisticians' seminar on "Strategic partnerships".

GE.16-02357(E)

I. Background

1. Statistics Canada presented an in-depth review of strategic partnerships with the information industry to the October 2015 meeting of the Bureau of the Conference of European Statisticians (CES) (ECE/CES/2016/5).

2. The CES Bureau commented positively on the review, citing its comprehensive coverage of partnerships, its use of examples and its identification of issues and challenges to be faced when forming partnerships. The CES Bureau suggested that the following topics to be discussed at the 2016 CES Seminar on Strategic Partnerships:

- How might partnerships benefit both the partners and society?
- How might partnerships work at the international level, including the Global Partnership for Sustainable Development?
- What is the role of Open Data in informal partnerships?
- What are the legal issues, and what can statistical offices do about them?
- Who are the other producers of data outside official statistics and what is meant by the 'information industry'?

3. The proposed aim of the seminar will be to:

"Determine the requirements of National Statistical Offices (NSOs) and other organizations for sharing information related to innovative strategic partnerships with the information industry, and determine the best means to do so."

4. This paper provides input to the seminar, according to that aim:

- What are the requirements for information industry partnerships in innovative statistical production?
- Who are the potential (international) information industry partners for innovative statistical production?
- What knowledge and information might be shared in a partnership for innovation in statistical production?

II. Building maturity in partnerships

5. In 2003, a seminar micro-data access was organised at Conference of European Statisticians plenary session[1]. Julia Lane gave a keynote speech with the following arguments:

- Improving micro-data access improves analysis, which improves the evidence for policy decisions, which in turn improves lives.
- The benefits to the research community and the scientific validity of research conclusions are substantial.
- There are also benefits to NSOs of providing micro-data access that can outweigh the risks and costs.
- A partnership approach between NSOs and the research community can maximise the benefits and minimise the risks and costs.

[1] www.unece.org/fileadmin/DAM/stats/documents/ces/2003/crp.2.e.pdf

6. In 2003, the relationship between NSOs and the research community was not business as usual. Typically, requests for access to confidential and unpublished data from the research community were treated by statistical offices as exceptions to a non-disclosure rule; as a non-core activity of a statistical office; and as risk with no benefit. Researchers were treated little differently to an "intruder" – indeed, many statistical offices used "intruder scenarios" to decide whether to grant exceptional access to their data for research purposes.

7. Julia Lane invited us to think differently. She described the NSOs and the research community as partners in a common aim. She suggested that through normalising accreditation of researchers and research projects and engaging with the research findings, NSOs can minimise costs, learn more about their data, and contribute to important insights that they do not have the capacity to investigate by themselves.

8. The 2003 CES seminar led to the development of the UNECE 2007 publication *Managing Statistical Confidentiality and Microdata Access*. This work brought together the perspective of the research community and the perspective of the official statistics community, and highlighted what was shared in common. The work has continued in the European Statistical System (ESS), OECD, and in national law, policy and practice.

9. The outcome is very substantial progress in the official statistics/research institution partnership. Today partnerships between NSOs and research institutions are business as usual. To a great extent the benefits predicted have been realised, the costs minimised, and the risks avoided. This component of NSO business has moved through the three stages of maturity, from 'naïve', through 'heroic' exploration, to an 'established' maturity. It has done so by realising that change was necessary; by challenging the culture that inhibited researcher access to micro-data; by sharing the perspectives of the partners; by building the legal and policy framework to enable change; and then by getting on with the job.

10. The example of improvements in partnerships for research access to micro-data of the last 10 to 15 years provides us with a model for developing our partnerships with the information industry, the subject of this CES seminar.

11. First, statisticians must recognise that change is necessary and that mature partnerships are needed.

- We should acknowledge that statistical offices cannot meet the global data challenge alone.
- We should recognise that the maturity of NSO partnerships with the information industry, formed for the purpose of innovation in statistical production, is today typically "naïve".
- We have some examples of heroic exploration. Examples of path-finding initiatives at seminars and workshops are a positive sign that we are moving from the naïve to the heroic stage of maturity.
- Being in the heroic stage is difficult. The costs and risks (and benefits) are experienced, but are managed only as exceptions. Knowledge and skill transfer is at a personal or project level, and vulnerable. Heroic statisticians are sure of their aims, but will be nervous and uncertain that they are doing the right thing which could inhibit their innovation. Moving through the heroic stage as quickly as possible is important.
- Mature partnerships are business as usual. The costs, risks and benefits are managed according to the usual business delivery of the office. Knowledge and skill transfer is at a programme or institutional level and assured. Under familiar governance structures our statisticians can be confident, and with confidence comes innovation.

12. Second, we must examine our culture, especially our inhibitions.

- Is there a barrier to partnerships with the information industry that is cultural? Are we worried about independence and dependency on others? Are we worried about public perception? Are we suspicious of the 'real' motivation of our potential partners, especially if they are commercially motivated?
- We have spent many years establishing that an NSO is different, that it needs its own law, independence, governance, stable budget, and policy role. We have been successful, and the Fundamental Principles for Official Statistics do indeed tell us we are special and different. Now this work is done and we have this protection, we can engage with others without fear of losing our important place in the national constitution.

13. Third, we need to review and revise our policy and legal framework. Our data sharing law, our information assurance and security policies, our governance and audit controls, our contracts standards – all these may need change. As an exceptional (heroic) activity, we may be tempted to make special arrangements, such as one-off contracts and memorandums of understanding about data handling. However, as a mature activity, partnerships with the information industry should be established within the standard laws, policies and governance of a statistical office – even if they have to adapt accordingly.

14. Fourth, we need to establish what must be done and do it. Changes in culture, policy, law, and practice will almost certainly need Project, Programme and Portfolio management to deliver the change; in particular where there are multinational components to the partnerships to be built. This requires clarity about business requirements and purpose, about benefits, and about contributions, which is the subject of the next section.

III. Why must NSOs form partnerships with the information industry in order to innovate in global data?

15. The Global Partnership for Sustainable Development Data exists to "support data-driven decisions with more open, new, and useable data[2]". It positions Official Statistics Systems as the building block for understanding development, and the partnerships with NSOs as the way in which data can "achieve higher quality, be more trustworthy, more understood, and open". The public service remit of official statistics (often expressed as 'better statistics, better decisions and better lives') is clearly aligned to the Global Partnership aims.

16. Our potential partners in the information industry are also aligned. Increasingly, with globalisation of information industries comes corporate social responsibility – a desire to improve lives and be seen to improve lives beyond the narrow impact of a particular commercial service or product. Perhaps this is not so new – Andrew Carnegie, Henry Wellcome, and Bill and Melinda Gates are examples of global business establishing permanent structures to deliver social benefits much wider in scope than the function of the original business. Carnegie's railroads, Wellcome's medicines, and Gates' computer technology all improved lives, but on a narrower basis than the foundations their legacy has created.

17. The Global Partnership is neither the only relevant partnership nor the only relevant expression of purpose. But it does express some expectations and key benefits for all

[2] www.data4sdgs.org

partnerships with the information industry for the purpose of innovation in official statistics. Usefully, it expresses that there are *societal benefits* in all our interests, *benefits shared* by the partnership, and *benefits to one or the other* of the partners.

A. Benefits to society

18. In the context of the Global Partnership for Sustainable Development Data, the benefits to society are the primary driver of a partnership. Understanding society to enable informed decisions to improve lives must be the ultimate shared aim of both official statistics and the information industry partners.

B. Benefits experienced by both partners

19. A partnership can deliver some benefits that both partners experience. Each partner will have separate and independent visions, values, strategies, and objectives. A successful partnership will be consistent with both sets of visions and values, will be referenced in both strategies, and will assist the delivery of each partners' objectives. It is here that the first tensions in cultural compatibility may be experienced when potential partnerships are explored, and the first changes in the policy frameworks may be required.

C. Benefits experienced by only one partner

20. It is to be expected that some benefits are experienced by only one or the other of the partners. In a contract for services this is recognised by a balancing payment. In a partnership between an NSO and an information industry organization the ideal would be to seek a net balance in partner benefits, to minimise any balancing payments.

D. Contributions expected of NSOs

21. Statistical offices have a public benefit remit, and may be expected to be culturally aligned to the *benefits to society* aims. In a partnership, the NSO can be expected to engage with the ministries, parliaments, and international representative institutions to formulate the new statistical requirement. The NSO partner can bring public confidence to the products of the partnership, and regulatory controls over the professional independence, quality, and scientific validity of the data.

E. Contributions expected of the information industry partner.

22. It may be expected that the information industry partner brings experience and expertise in alternative sources of data. Expertise in data on the interaction with customers, events/transactions data, and 'big data' is more likely to be found in the information industry partner. The information industry partner probably has the necessary technology and processing power to use these data and to integrate them with the spine of data held by the NSO partner, especially because their success as a business will be founded upon extracting the value of their data assets.

F. Contributions that can be expected of both partners.

23. Existing official statistics can be expected to be the bedrock of insights (civil registration, environmental accounts, trade, household expenditure, and census data, for example). To innovate beyond these bedrock statistics, the NSO partner should expect to provide its information industry partner with privileged access to upstream stages of this processing, in order to investigate how other sources of data and other presentations of data can be built in to innovative production. Similarly, the information industry partner can be expected to provide privileged access to their unpublished upstream versions of their data sources and the technical capability for analysing them. Both partners have a strong culture of denial of access to these unpublished, often confidential, upstream sources, and a major contribution of both partners will be the conditional agreement of access to these data as is necessary for achieving the societal benefits of better data.

1. Open questions about the formation of partnerships with the information industry

- Can we agree that partnership with the information industry is now a pre-requisite for innovation in official statistics?
- Can we expect official statistics partners (public sector) and information industry partners (typically private sector or third sector) to be seeking the same shared benefits for society and for the partnership? Is this independent self-interest coming into alignment for a while? Or is this a genuine and permanent common purpose?

IV. Potential international information industry partners

24. Official statistics pioneers in building partnerships with the information industry have worked with, broadly speaking, two types of partners. NSOs have sought partners who appear to them to be wells of data, and the partnership is about exploiting those data sources for the purpose of improving official statistics. For example, statistical offices have approached supermarkets and energy companies to obtain access to their loyalty card records and the smart meter data of their customers. NSOs have also sought partners who have expertise in managing alternative data sources such as management information, customer records, and big data. These partners are valuable because of the help they can give the NSO with bringing these sources into statistical production. Engagement with these partners may come from participation in research projects and initiatives such as Horizon 2020. In other words, these partnerships are often asymmetrical, and are mostly about the benefits that may accrue to the official statistics partner. Of course, a societal benefit comes from those better official statistics, yet it may be fair to say that these partnerships do not benefit the industry partner equally, but do impose costs and risks. This may explain why the approaches NSOs make often fail, not least because the industry partner may expect the asymmetry to be addressed through payments for risk compensation which NSOs may be unwilling or unable to pay.

25. The search for an information industry partner should start from a different premise. NSOs should identify partners that share the ambition of ‘better statistics, better decisions, better lives’, and who can join in a common programme of work where there is as much symmetry as possible in the contributions made and the benefits realised. A very good and recent example is the formation in October 2014 of a Public Private Partnership[3] between the European Commission and the Big Data Value Association (BDVA)[4]. The BDVA is a

[3] http://europa.eu/rapid/press-release_MEMO-14-583_en.htm
[4] http://www.bdva.eu/

self-financed not-for-profit membership organization, consisting of a number of organizations with valuable data sources, data skills – and a shared ambition to exploit those data for the mutual benefit of all. The partnership agreement commits the Commission to 0.5Billion Euro of investment into developing Big Data resources, which will be matched with 2 Billion Euro of investment from the BDVA members – a very substantial information industry partnership.

1. Open questions

- Is access to information industry partners unequally distributed, internationally? If so, how should this be addressed?
- Is the Public Private Partnership model the best one for these purposes?

V. Defining the "Information Industry"

26. The "In-depth review of strategic partnerships with the information industry" used the definition of Wikipedia for the "information industry", which includes six categories: (i) producers and sellers of information, (ii) information-processing services, (iii) dissemination services, (iv) manufacturers of information-processing devices, (v) research-intensive industries, and (vi) providers of infrastructure for information production and sophisticated decision making. Outreach and engagement activities that can lead to specific partnerships are also included in this review.

27. As an alternative, and / or supplementary approach, the definition for the ICT sector could also be considered. The ICT sector covers industries active in Information and Communication Technologies. They are defined by NACE/ISICs.

28. The ICT sector definition based on NACE rev. 2 classification is as follows:

29. The ICT sector consists of all enterprises/units (including both natural and legal persons) which principal activity (principal activity contributes 50 and more percent to the value added) belongs to following divisions and groups (classes) of NACE rev. 2 classification:

A. ICT manufacturing industries:

- Manufacture of electronic components and boards – group 26.1
- Manufacture of computers and peripheral equipment – group 26.2
- Manufacture of communication equipment – group 26.3
- Manufacture of consumer electronics and related media (groups 26.4, 26.8):
    - Manufacture of consumer electronics – group 26.4
    - Manufacture of magnetic and optical media – group 26.8

B. ICT trade industries:

- Wholesale of information and communication equipment – group 46.5

C. ICT services industries:

- Telecommunications – division 61:
  - Wired telecommunications activities – group 61.1
  - Wireless telecommunications activities – group 61.2
  - Other telecommunication activities (groups 61.3, 61.9):
    - Satellite telecommunications activities – group 61.3
    - Other telecommunications activities – group 61.9
- ICT services industries (division 62; groups 58.2, 63.1, 95.1):
  - Software publishing and IT service activities (division 62, group 58.2):
    - Software publishing – group 58.2
    - Computer programming, consultancy and related activities – division 62

      - class 62.01 – Computer programming activities

      - class 62.02 – Computer consultancy activities

      - class 62.03 – Computer facilities management activities

      - class 62.09 – Other information technology and computer service activities
  - Data processing, hosting and related activities; web portals – group 63.1
  - Repair of computers and communication equipment – group 95.1

30. Most of the industries within the ICT sector would qualify as potential partners for further developing the system of official statistics. However, we can also observe businesses outside the ICT sector that invest heavily in data (collection, processing, analysis and dissemination) services. This is due to the development of dematerialisation of production processes. We can observe more and more businesses that do not produce any goods but control the entire production process and sell the products under their name. Other big businesses in manufacturing are more and more developing to service providers, such as the big car manufacturers who heavily invest into software for intelligent car management and try to build services on top of the collected data to be able to offer goods (cars) and services because expectations for future growth are associated with services.

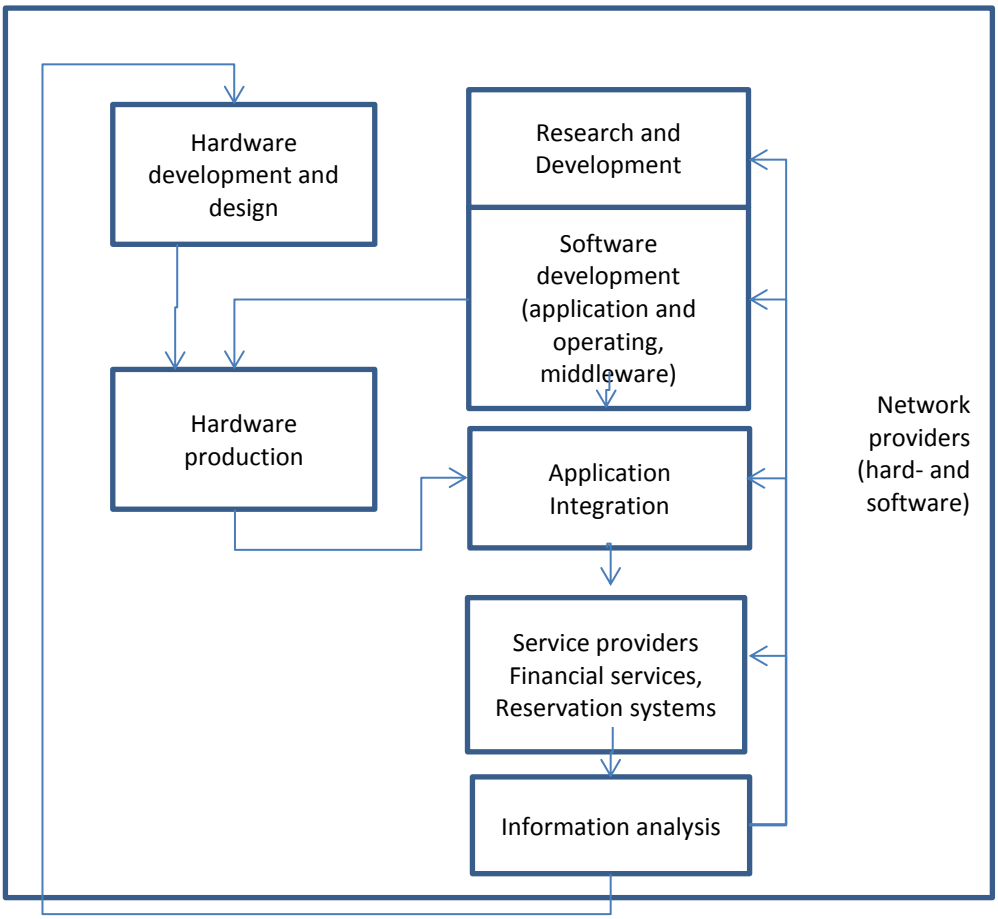
31. The World Economic Forum recently published an article on the seven technologies that drive the fourth industrial revolution (http://www.weforum.org/agenda/2016/01/a-brief-guide-to-the-technologies-changing-world). Businesses associated to these technologies would be very suitable as partners as they develop the ecosystem of future markets. In addition, the referred developments are associated with generating huge amounts of data that can be secondarily used for statistical purposes. Development of intelligent systems is necessary to filter aggregate data and transform data into information for taking decisions by autonomous systems or by humans. Statistics, as one purpose, could be integrated in those systems. The technologies are:

- computing capabilities, storage and access
- relates to computing capacity of mobile devices and access to almost unlimited storage space via the internet.
- Big Data

- relates to creation of massive amounts of data that can analysed for improving decision making and advancing artificial intelligence.
- Digitization of matter
- relates to 3D printing that will bring design and manufacturing and enable personalised products online.
- Digital health
- relates to digitization of health information that will enable new areas of applications related to people's health and development of services related to the domain.
- Internet of things
- relates to deployment of trillion of sensors during the coming decade will enable automation of many processes in professional and private life.
- Wearable internet
- relates to implantable and wearable devices directly or indirectly connected to the internet.
- Blockchain
- relates to verifying and assuring integrity of transactions.

32. Another approach for identifying businesses that would be most suitable for partnerships in the future is to analyse service delivery process. As noted above, we observe an increasing tendency of dematerialisation (digitization of matter) and the growth of service providing companies. Information on products and clients is more and more collected via sensors (internet of things, wearable internet) and other means of analysing behaviour of persons. The information is analysed and used to define new services (where the physical good comes with the service) or improve existing services. Businesses that run those integration and services platforms or which develop soft- and hardware for these platforms are well suited candidates for partnerships. In addition, researchers in universities or in research departments of private companies as well as start-ups often lead the technological development, which makes them as well good candidates for partnerships.

Figure 1
Service delivery process



33. A number of these businesses operate internationally and, while there are national specificities, in most cases models of collaboration can be replicated from country to country and can also be extended to truly multi-country partnerships. In addition there is a tendency that one service provider gains a considerable share of the market, such as Google for the search engines, Facebook for social media, or mobile communication is offered by a few network operators by country. It would be useful to approach these multinational actors in a coordinated way in order to maximise benefits of partnerships across countries and to establish common principles of collaboration.

D. Types of partnerships for innovative statistical production

34. The typology derived from the UNECE survey on partnerships in Big Data and official statistics distinguishes between the following partnerships:

- Data provider
- Data consumer / aggregator
- Design partner
- Technology partner
- Analytical partner

- others

35. In the guidelines document the above partner roles were condensed to four groups.

- Provider
- Design and Analysis
- Technology
- Users, coordination and communication

E. Partnerships by type of partner

36. The following text describes certain dominant characteristics which need to be considered when preparing the establishment of a partnership between NSOs and third parties (mainly industry actors). While current practice is often limited to single types of partnerships (i.e. with one type of partner) it would be desirable for NSOs to explore multi-partner partnerships in the near future. Another important element which should be considered is the dynamic aspect of the big data ecosystem in terms of near future changes and their impact on the respective roles of the various partners.

F. Provider

37. The provider partnership relates to data acquisition from data sources/providers. This includes:

- New and alternative providers and sources or new sources from existing data providers
- Data providers undertaking upstream processing so that statistical organizations do not need to invest in the processing infrastructure, capability or storage.

38. Currently, the main challenge for big data projects is getting access to data, while managing privacy and confidentiality. Business models of data providers could conflict with getting free access to data and free dissemination of statistics derived from big data. Statistical organizations can collaborate and develop strategies towards data providers (for example mobile network operators) to address these challenges. These types of partnerships represent the most frequent cases amongst those which were reported in the UNECE survey.

G. Design and Analysis

39. There is strong interaction between design and analysis. Statistical organizations can work with partners to conceive research questions and to co-design and develop projects that address the relevant statistical and development challenges.

40. Statistical organizations desire to produce high quality data. This might affect how partners approach the design, and could require new methods and different approaches. For example timeliness could be increased by using Big Data on the expense of other quality dimensions. Quality requirements might be very different compared to survey data.

41. Engineers, data scientists or researchers can collaborate with statistical organizations on specific projects.

42. Statistical organizations can provide standards and methodology and other organizations provide analytical capacity and modelling. This is the second most popular type of partnership found in the UNECE survey.

H. Technology

43. Access to the best tools for data processing, data mining, real-time analytics, storage, computing, and data visualization is essential to our data scientists' ability to successfully tackle research questions. The private sector can provide analytics software that statistical organizations can use to sift through vast quantities of data to tell its hidden story, look for emerging trends, etc. They can provide modern software applications that can leverage and harness Big Data in order to gain new insights.

44. It is difficult to choose partners in Big Data technology. At the moment the market is not mature and is still evolving continuously. There also isn't much standardisation. There is a lot of software in the open-source category so potential partners are in both the private sector and open source communities, which brings new challenges in service arrangements.

45. The Big Data project of the European Statistical System (ESS) identified a number of big data sources that could be subjects to pilot projects in the future. Private businesses related to these sources could be candidates for "provider partnerships".

46. Big data sources for pilot projects include:

- Detailed transaction level data from telecom operators or operators of alternative voice or data communication networks
- Sensor data that can be related to individual persons or households:
    - Data from personal communication devices
    - Data from wearables, e.g. smart watches, heart rate monitors, etc.
    - Data from smart electricity consumption meters (smart meters)
- Non-personal sensor data:
    - Road traffic loops
    - Remote sensing data including satellite image data, data from unmanned aerial vehicles (UAV), etc.
- Data obtained from the internet including:
    - Social media data with or without geolocation information
    - Web-scraped data from company websites, e-commerce websites, job vacancy websites or real estate agencies' websites
    - Query and click-out data from internet searches
- Financial transaction data (credit cards, debit cards, online payment systems)
- Personal health data such as Electronic Health Records, including data from health registries, laboratories, donor programs, etc.
- Electronic reservation systems data - e.g. data from flight or hotel booking systems
- Cash register data, e.g. from supermarkets

VI. What knowledge and information might be shared in a partnership for innovation in statistical production?

47. A successful partnership striving for innovation in the statistical production process is dependent on addressing a mutual interest for cooperation by the NSO and the private partner. However, being aware and making use of each partner's comparative advantage during the course of the partnership is of high importance for effective and successful results.

48. Among private partners one needs to distinguish between *private data owners* (e.g. mobile phone companies) and *information providers* (e.g. Google) regarding their comparative advantage. Usually the purpose of data collection by a *private data owner* differs from the intended statistical use, which opens a window for cooperation as the NSO is no competitor, but will add an application to the already collected data. The comparative advantage of private partners, who are data owners, is first of all the ownership of the data of interest, but also profound knowledge of the collected data, which enables efficient utilization of the data. Moreover, having existing consumer relations with the data providers can be a valuable factor for further requests or the possibility to match data with administrative data sources.

49. The comparative advantage of a private *information provider* such as a search engine supplier is the ownership of large volumes of data which can bear the possibility to be used as data sources for official statistics. The difference of *information providers* to simple *private data owners* is that their collection of data is not only a by-product, but serves the purpose to provide deeper and more sophisticated information to their clients by combining different sources of information. As their main purpose is to provide information to their clients they also have a high level of technical skills and sophisticated ways of data visualisation and presentation. This includes the capacity to cope with semi structured data of high volume and to provide information in a very customer friendly way on new technological communication devices. Their focus is on timely of data collected as this what people are interested in. They put a lot of effort in investigating their users' preferences, the wide range and high number of users gives those information providers a high potential as dissemination multipliers. Other than NSOs, information providers have high investment capacities, which enable the development of innovative projects (e.g. the development of an app) or large-scale projects.

50. The NSO is able to contribute through its high level of methodological expert knowledge and long-term experience as data producer and analyst. Data produced by official statistics usually describe the populations of interest without severe coverage problems. Harmonized concepts are stable over time and enable users to compare the figures over time and across countries. Experts at the NSOs have profound knowledge of the relevant statistical processes and are well aware of other relevant data sources and how to overcome coherence problems when combining different data sources. This knowledge can help data owners to get to know the applicability of their data and to expand the use of their data in the future. In most cases a NSO is embedded in a kind of international environment like the United Nations (UN) or the European Statistical System (ESS). The need to comply with common rules of such systems for instance the European Code of Practice or the fundamental principles of the UN create trust of users in the products of NSOs. A private partner can gain form the trustworthiness of official statistics. One comparative advantage of NSOs is that they have access to administrative data – an information source not systematically open to private companies.

51. The partnership between data owners/information providers with NSOs can include:

- Skills and knowledge transfers between NSOs and private partners, respecting the other partner's comparative advantage;
- staff exchange:
- the access to skills development courses:
- the development of common data projects.

52. A successful partnership between private companies and an NSO is bound to some conditions. This includes the full respect of privacy rules by both partners, costs incurred on each side throughout a cooperation project will have to be carried by each partner. The cooperation cannot include privileged data access for the private partner; and data by the private partner are made available to the NSO free of charge.

VII. INEGI, Mexico – an example partnership to exploit Twitter data

A. Introduction

53. Information from Internet systems and from electronic devices connected to the network can contribute to the production of statistical and geographic information. This is why international organizations and national statistical offices from several countries, including Mexico's INEGI, are dabbling in practical applications of data science to resolve problems of Big Data. As part of INEGI studies in the field of subjective wellbeing, we decided to use Twitter as a source of big data to determine the mood of Mexican tweeters.

54. A multidisciplinary working group was formed with INEGI, and a group of national and foreign academic institutions– INFOTEC, Centro Geo, University of Pennsylvania and the Tec Milenio University.

B. Collection of data from Twitter for statistical purposes.

55. Twitter is a social network in which users type short texts, up to 140 characters long, which are publicly available, which means that anyone can read them. Additionally, Twitter has the option of "geotagging" tweets; that is, labelling each tweet with the geographical coordinates of location at the time of publication. The Analysis of Mexican Twitter's Mood focused on these geo-referenced tweets, which can be downloaded through geographic filters regardless of the topic the tweeter dealt with.

56. Using publicly available Twitter apps, INEGI has collected and geo-referenced tweets within the Mexico's territory, southern USA, and the northern part of Central America.

C. Geocoding tweets

57. In order to generate statistics at the state level, each collected tweet was assigned the state and municipality geostatistical codes belonging to the place of publication. In other words, this geographic analysis classifies tweets according to the state and municipality they were published from.

D. Generating the manually labelled subset

58. In order to generate statistics regarding the mood of Mexican tweeters, it was necessary to classify each tweet according to the emotional charge that identifies the mood the tweeter was in when he wrote the tweet. Doing this by hand would be a monumental task since our database already contains several million messages, so "Machine Learning" techniques are used.

59. First the manual sorting of a subset of tweets in which a label is assigned according to the emotional charge of each tweet is required. The label assigned to each twit is defined as positive, negative or neutral.

60. To generate this subset of tagged tweets, collaboration with Tec Milenio University was established. This made possible to count with more than 5000 students who manually labelled thousands of tweets. In this exercise, every tweet was presented more than once to one or more students so that a single tweet could be labelled several times, in search of a consensus on the label.

61. Subsequently, manually tagged tweets underwent as cleaning process in which we sought to reduce uncertainty by entropy-based techniques in order to reduce clutter on ratings. Thus, we identified and removed the tweets of inconsistent taggers, contradictions and repetitions were discarded, and those tweets with greater consensus on the label were identified as well as those from students who showed greater consistency.

E. Development and training automatic classifiers

62. Innovative statistical learning algorithms were developed using artificial intelligence techniques by researchers from INFOTEC and Centro Geo. These algorithms were integrated forming an ensemble.

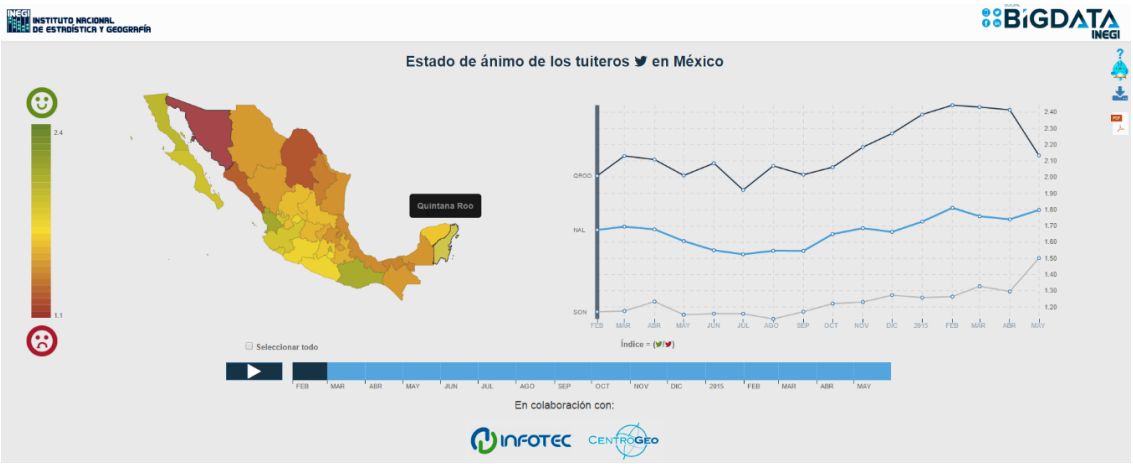
63. The ensemble, based on the accuracy of classification of individual algorithms, makes the most of each individual algorithm achieving around 80% accuracy in labelling of a different subgroup of manually labelled tweets.

F. Massive classification of tweets

64. Using the trained ensemble of algorithms, all remaining previously normalized twits were processed, resulting in a database of twits with a new attribute indicating the emotional charge of each twit.

G. Tool for displaying statistical mood of tweeters in Mexico

65. Finally, we developed a visualization tool that takes the result of automated classification of 63 million tweets to represent the spirit of the tweeters in Mexico, showing statewide breakdowns per month. An index representing the ratio between positive and negative twits was developed. This index is represented by means of both graph and maps.



http://www.inegi.org.mx/inegi/contenidos/investigacion/Experimentales/animotuitero/default.aspx

H. Other statistical exercises using Twitter

66. Geo-referenced twits have proven useful also for the study of mobility patterns. For instance, people who continually tweet from the same state (or municipality) within Mexico are easily identifiable as short-term visitors to a different state, when the new location of their Tweeter activity changes back to the original state. We have tried this for visitors to the states of Guanajuato and Puebla over a well identified holiday period in 2014, and managed to identify an increase in Tweeter postings from visitors when compared with non-holiday weekends. We have not yet measured the mood of holidaymakers. We think that this will help explain, at least partially, why states with important tourist activity throughout the year consistently rate higher (painted in green in the map) in Tweeter mood than all the rest. Other mobility studies can be performed in a similar fashion. For instance, short-term circular border crossings; or daily trips to and from school or workplace, with the routes depicted, on top of origin and destination points; or paths followed over the Mexican territory by Central American migrants on-route to the US.

67. We have also been asked to look into the feasibility of automatic identification of mental health conditions in adolescents and young adults by analysing their publications on social networks. To this end, we have included in the group researchers from the Mexican Institute of Psychiatry (IMP). Briefly, young diagnosed patients with an account will be invited to make their Tweeter id and postings available. Postings will then be tagged with one main diagnosed mental condition; i.e., they will be manually tagged as above. From this point on, the study may proceed in a similar fashion to the one about the tweeters mood. New postings will then be tagged as more or less likely to reflect a particular mental condition.

VIII. Conclusions

68. The example from INEGE, Mexico, illustrates the benefits of working with information industry partners. Derivation of mood states from 140 character Tweets, machine learning, entropy based techniques to reduce clutter, algorithms derived from artificial intelligence, and identification of mental health conditions – these abilities are not found in the permanent staff of many statistical offices, and without them a valuable source of data goes unexploited by the NSO owners of the bedrock data for global statistics.

69. This example is ‘heroic’, and there are others to exchange, and these can inspire us to move from this maturity stage to the point where working in this way is embedded in what NSOs do, and becomes our first choice when addressing the challenges of global data for better statistics, better decisions, and better lives.