Distr.
GENERAL

Working Paper
28 February 2014

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2014)
(Dublin, Ireland and Manila, Philippines 14-16 April 2014)

Topic (iii): Innovation

Big Data uses cases and implementation pilots at the OECD

Prepared by Jens Dossé, OECD

I. Introduction

1. Different Big Data needs, innovating on data sources and on the ways OECD collects and analyses data, are currently emerging at the OECD, and a special team has begun to identify and analyse use cases, potential solutions and implement pilot projects. Our approaches are inspired by work carried out previously by other organizations such as Eurostat and Statistics Netherlands who have already progressed significantly in domains such as traffic pattern analysis using mobile positioning devices, well-being trends through social media analysis, and internet crawlers for price statistics. We are looking forward to a potential collaboration on existing solutions with these organisations, and would be ready to share our own approaches and solutions with other organisations.

2. The currently identified use cases at the OECD are:

II. Real estate price

A. Use case description

3. The objective of this first use case in the Directorate for Public Governance and Territorial Development (GOV) is to collect real estate prices in real time by extracting information (e.g. location, features of products and price information) from real estate ads sites in main population centres. We have examples, particularly in the U.S., where agencies have a wealth of information concerning each real estate with information such as: garden, number of rooms, number of bathroom, information about neighbours, on the size of the house of the neighbours, their number of pieces, etc.

B. Proof of concept envisaged

4. The GOV Directorate has identified 8 key real estate websites in different countries such as England, France and Spain. Data collection would be through a search engine with semantic analysis and structured data collection capabilities. Following this, the available statistical tools allow analysing the data to produce aggregated indicators. The last step is to enrich GOV's Metropolitan database with the compiled indicators. This work will be greatly inspired by Statistics Netherlands' work on real estate internet prices.

III. Real-time traffic information

A. Use case description

5. The International Transport Forum (ITF) is aiming to collect a set of mobile data to study the traffic situation at given locations. The objective is to also get the evolution of the situation in real time.

B. Proof of concept envisaged

6. The main obstacle identified here is the use of private data. We have identified a start-up company specialized in collecting mobile data that uses a mechanism of voluntarism. Indeed, users volunteer to have their information collected. For that purpose, a chip is installed on their smartphones. However, the number of volunteers has to be significant, and this study should be done at international level by targeting users through different countries. Then a data analysis using existing statistical tools would allow producing aggregated indicators. This data could also be used by the Science, Technology and Industry Directorate (STI) in a study on the quality of the mobile network, and by the Directorate for Public Governance and Territorial Development (GOV) for a definition of metropolitan area, based on the intensity of transport.

IV. Political and social tensions

A. Use case description

7. In the context of the African Economic Outlook (and more specifically for the chapter on economic governance and policy), the Development Center (DEV) built indicators that indicate the level of political and social tension in 53 African countries. To do this, during 2 months per year, 2 persons read thousands of articles in the press, identify keywords (strike, demonstration, violence policies, kidnapping, etc.), and then assess the severity of the events. For example, the severity of a demonstration is considered as low, medium, or high depending on the number of participants in the demonstration. DEV would like to automate this tedious reading exercise. An automated analysis would enable substantial time savings, allowing allocating more time to the study, analysis and interpretation of these situations.

B. Proof of concept envisaged

8. Textual data analysis might be the answer to this need. It corresponds to the multidimensional descriptive analysis of text. Such a study can be performed on a number of domains such as analyses of speeches, press articles, socio-economic surveys. Textual data analysis does not relate only to the quantitative analysis but also to the analysis of the content, text mining (text analytics), information retrieval and computational linguistics (synonymy, homography, lemmatization).

9. Analysis of textual data therefore represents a relatively significant interest in the use of Big Data. It permits to classify the different sources depending on their origin, dates, etc. A study is therefore feasible on a set of articles by targeting key words in order to have an analysis on each country of the political situation as is the use case for DEV division. An approach would be to implement code dictionaries for the identification of political events, actors, quantities (e.g. "a few thousand", "about 30000", "hundreds") and

duration (e.g. “about a week”, “for 3 days”). During the text analysis, and possibly semantic analysis, it would be important to identify also the cause of the event, e.g. through query optimisation.

V. Confidential micro data

A. Use case description

10. Directorate for Employment, Labour and Social Affairs (ELS) works primarily with voluminous micro data (on individuals) collected through companies and household surveys. More recently they aim at a collection from administrative registries of national authorities and databases of internet providers, which has the difficulty of legal complexity. Related access restrictions prevent an optimal analysis.

11. The objective of this use case is to get a widespread access to micro data and find solution to assure confidentiality.

B. Proof of concept envisaged

12. Potential solutions include the establishment of an international platform that would make the micro data available to organisations such as the OECD without the possibility to store this data, or the implementation of filters in order to block access a certain level of detail (disclosure control on Big Data).

VI. Job demand

A. Use case description

13. Some countries such as the USA, Canada, and New Zealand use the Internet in order to carry out statistics on demand for employment. Already today to gauge the job supply, the Directorate for Employment, Labour and Social Affairs (ELS) crawls the internet and aggregates the information found to calculate advanced indicators. It would be interesting to take this model and apply it to the job demand. Thus, the solution would be to deconstruct this model to understand the functioning and then apply it to the demand for employment, particularly by retrieving the number of respondents to the job offers. Then either the information could be picked up from sites such as “Monster” or bodies like ANPE accept that we can query it. This poses a legal problem that would be solved with the anonymisation of data.

14. The project concerning job supply, vacancies and e-skills also interests the Directorate for Science, Technology and Industry (STI). They want to base their studies on internet sources such as the job sites Monster and Apec. The information would be useful to conduct an analysis of the situation of employment specifically on the internet.

B. Proof of concept envisaged

15. A crawler could be implemented. The cost of a specific development (e.g. with R or SAS) is to be compared with the costs of accessing databases of data-collecting organizations.

VII. Laboriousness indicators

A. Use case description

16. This use case also in the Directorate for Employment, Labour and Social Affairs (ELS) concerns the scoring of texts on labour permitting to measure the degree of laboriousness of the various facets of the

world of work in the different countries. The current study is based on surveys followed by a lengthy verification by different teams. So far, 40% of the answers contain incorrect responses. The main problem here is that this type of investigation from analysis of legal texts is politically very sensitive as many policy decisions are made as a result of the findings. This therefore raises the complication of the reliability of the input, and it seems quite difficult to perform a crawling on the internet with data for which reliability cannot always be verified. The current scope would therefore be on civil codes of the various countries studied, replacing the questionnaires.

17. The goal is to facilitate research, reading and consolidation of information through a tool for semantic analysis, with the implementation of a legal dictionary helping to automate this work. Big Data might therefore be the solution to ELS due to the fact that the studies carried out by the Directorate requires a large and growing amount of data while in return the number of related queries decreases.

B. Proof of concept envisaged

18. A proof-of-concept should demonstrate that a textual data analysis tool can meet the complex requirements.

VIII. Internet connection quality

A. Use case description

19. As explained earlier, the International Transport Forum (ITF) wants to assess the traffic flow through chips embedded in smartphones, The Directorate for Science, Technology and Industry (STI) would like to reuse this process in order to assess the quality of the signal, the quality of the connections, and how the smartphone applications are used.

20. A data study on the use of the internet in order to analyse the quality of internet connections, as well as an analysis of the quality of the infrastructure are to be carried out. However, this project being very specific, it requires intense research while knowing that the usefulness of this research cannot be determined before it is completed.

B. Proof of concept envisaged

21. Crawler and text analysis tools should be challenged on this use case. Alternatively, specific developments might be necessary or more cost-effective.

IX. "Machine to Machine" communication

A. Use case description

22. Within the project "Machine to Machine" another longer term goal of STI is also to study the use of the various internet communications tools. Today the users are humans; however forecasts within the next 5 years see a majority use of these different technologies by machines to communicate with each other.

B. Proof of concept envisaged

23. Again, crawler, text analysis tools or specific developments should be employed on this use case.

X. Internet language ranking

A. Use case description

24. A study on the ranking of the languages used on the internet is also possible. What are the most used languages on the internet? Already knowing that English is the leading language on the internet, the language ranking would serve to identify the following languages

B. Proof of concept envisaged

25. The approach is thus, by defining a panel of generic and more specific sites, to count the appearance of each language. A question would be whether to limit the number of languages included in the study.

XI. Quality and usefulness of internet security

A. Use case description

26. Looking at the protection of intellectual property on the internet, are the means applied in the past still valid? Many tools are used to defend intellectual property on the web. Surprisingly, in recent years more and more artists put their works for free download on the internet. This raises questions on usefulness, quality and new ways of security measures.

XII. Prices of traded goods and services, and foreign investments

A. Use case description

27. The Trade and Agriculture Directorate (TAD) is interested in the issues with collecting voluminous data that are more recent than survey data. The data targeted are trade prices and transport movements as already envisaged by ITF with transport data. The Trade sector is looking for the prices of goods and services that are not in the TRADE database and is thus interested in the crawling on merchant sites. It also looks for data on investments between countries and multinationals.

28. The objective is to limit speculation on raw materials so that some populations are not deprived of food. This is linked to the AMIS project which consists of collecting and reading many articles and gathering raw material prices, with the aim of anticipating any irregularities that could lead to tensions in the raw material market. It could be observed in the past that tensions and crises could be better prevented if the price information available to various international organizations would be pooled together and visualized to see the evolution of raw material prices in real time and early detect abnormalities such as droughts. This would allow activating special measures to avoid speculation.

29. The Directorate lacks information and data on traded services. The definition of a basket of services in order to compare the price of services and the evolution of prices between countries is therefore necessary.

30. Finally, the TAD Directorate would like to try to get a better view on the amount of trade between the countries or large companies in terms of investment. Since the last economic crisis, it becomes necessary to have up-to-date data but sources are sometimes missing. The Big Data would allow getting exploratory data for trend analysis that would then be confirmed by specific studies that currently rely on surveys in t-2.

B. Proof of concept envisaged

31. Crawler and text analysis tools should be able to fulfil the requirements which for the first part have similarities to the Real estate prices use case.

XIII. World Input-Output Data (WIOD) and Weeds data etc.

A. Use case description

32. The Trade in Value-Add project (TIVA) project in the Trade and Agriculture Directorate (TAD) currently exploits old survey data and private data such as ORBIS. The first concern is the recency of the information that is no longer appropriate, and the second is the current difficulty to cross all the data. The use case is thus a about crawling, cleaning, reprocessing and requalifying data in free access from e.g. World Input-Output Database (WIOD) and Weeds databases.

33. The data is unstructured because using data from customs tariffs, trade agreement (qualitative work) and non-tariff measures. A data hub project GTAP has emerged to allow validating the data by a consortium of organizations. A remaining problem is that the OECD, as an international organization, cannot access a number of private data that are essential for the Directorate's work. The setup of access and sharing policies could quickly become necessary for an optimal use of Big Data.

34. TAD aims also to cross the TIVA data with information from the Services Trade Restrictiveness Index project (STRI).

B. Proof of concept envisaged

35. A crawler/scraper tool would be required.

XIV. Measuring the impact of surveys

A. Use case description

36. Paris 21 is the secretariat of a group of agencies including the OECD (represented by its Directorate for Cooperation and Development (DCD)) as well as many statistical institutions in developing countries. This group participates in the development of statistical analysis as well as the funding of statistics in those developing countries. The members of this group have set up an Accelerated Data Program to work with surveys and micro data. Via this channel, if the OECD has tools that can treat the Big Data, this will help meet the needs of these developing countries.

37. Paris 21 develops a study concerning the use of survey results in these developing countries. It performs an aggregation of Google search results on these about 3000 surveys. Then it searches all the articles where these surveys are cited or used. This will permit in term to identify the use of these surveys, their necessity (especially concerning their funding) and their relevance. Since this work would represent a huge manual effort, automation would be very useful, even more as it will have to be repeated and updated.

B. Proof of concept envisaged

38. Again, crawler and text analysis tools might be able to meet the need.