WP.33
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (v): Changing organizational cultures

Availability of data: from scarcity to profusion

Invited Paper

Prepared by Jean-Pierre Kent, Statistics Netherlands[1]

I. Introduction

1. The production of official statistics is a mature process that reflects the efforts made to develop it and bears witness to the context in which this development took place. This process uses quality criteria that are in harmony with the needs of users in this context. As long as this context is stable, user needs remain stable too, and it is reasonable to keep applying the same process and to focus on perfecting it.

2. This context, however, is changing at high pace, and it is prudent to consider the consequences of this evolution for our users and their needs, for the quality criteria of our products, and finally for our production process itself.

3. This paper compares the context of statistic production of the 21st century with that of the 20th century and examines the signs of change in user needs and user expectation. It also draws a parallel with two industries that were confronted with a similar phenomenon during the 20th century, and extracts lessons that are becoming relevant for the production of statistics.

II. From scarcity to profusion

4. The 20th century confronted statisticians with a tough challenge: they had to produce data about general phenomena like economy and demography, while raw materials (individual data about persons and businesses) were not readily available. Data collection methodology has become part of our toolbox, and we tend to forget the efforts that were deployed to develop it. It was a tough challenge nevertheless, and we still like to think that statistics will always remain a prerogative of specialists.

5. Today, however, data are omnipresent. The challenge is no longer to extract data from people or from enterprises, but to select and use data that are readily available. Some schools have already started to teach children how to access Google Trends, how to select relevant data, and how to produce meaningful graphs and summaries. The upcoming generation will consider statistical skills to be common knowledge. Just like today's youth produce their dissertations in camera-ready form (which was unconceivable only 15 years ago), tomorrow's youth will be producing their own statistics (which today's statisticians still find hard to admit).

6. The signs of this change are clear. The most outstanding example is the GPI (Google Price Index). By using a fully automated process to produce a price index based exclusively on Internet data,

[1] The opinions expressed in this paper are under the responsibility of the author, and do not necessarily reflect the policy of Statistics Netherlands.

Google show the feasibility of cheap and timely statistics [3]. Their figures do not meet the quality standards of NSIs, however. Google's chief economist, Hal Varian, indicates that this has to do with the representativity of data available through the Internet.

7. Cheap and timely products always have deep impact on users' quality criteria: they cause price and time to market to become the dominant criteria. Statistics will be no exception. We are no longer able to state that quality comes at a price and takes time. We need to assess the impact of these developments on users' quality criteria and adapt our production accordingly. Quality becomes negotiable, but needs to be objectively evaluated and communicated explicitly.

8. Access to private data is clearly an advantage for NSIs. The challenge is to combine this trump with the chances offered by the masses of data available on line. This challenge needs to be met now. What is at stake is the relevance of official statistics. Our relevance has already started to fade. This can be seen in the fact that self-made graphics and references to internet-based statistics appear in economic or demographic articles published in newspapers, where we were used to see only references to publications of NSIs, central banks and other official instances. In my view the most outstanding example of our fading relevance is the European Commission's project "FOC", aiming at the production of early crisis indicators: it involves the European Central Bank, several European universities and Yahoo [4] [5]. A few years ago no one would have dreamt of conducting such a project without the participation of Eurostat, and maybe Yahoo would not have been considered to be an adequate partner.

9. Use of Internet data for the production of official statistics, however, is not as simple as it might seem. Our statistical process is characterised by a number of best practices that were developed in the context of data scarcity of the 20th century. It is not obvious that these practices should also be adequate in today's context of data profusion.

III. Best practices

10. Let us turn our attention to a few of these best practices. In the following paragraphs we will examine the consequences of the new situation for the following:

- Primary data collection
- Purpose orientation
- Data editing
- Trend continuity
- Production by subject matter specialists

A. Primary data collection

11. Until the mid nineties it was taken for granted that the only way to collect input for our statistical production was to get people to fill in forms or to answer interview questions. Then we gradually included the use of official registrations. We determined which registrations would be able to fit our needs, and strived to process and edit them to replace or supplement survey data. Today there are still statisticians who reluctantly use registrations, underlining the fact that administrative data are less suited for the production of statistics.

12. Internet companies, however, have demonstrated that something very interesting is happening. We are in a changeover from a society with little or no data available to one that has an abundance of data. We see other parties making statistics that are similar to ours but much cheaper and much faster (e.g. Google [3]) and on an almost global scale. The question is no longer whether available data are adequate for the production of our traditional statistics, but whether our traditional approach to statistics is adequate for processing the data that are readily available.

B. Purpose orientation

13. Traditionally the output of NSIs has been determined by the demands of their respective governments and other organisations. The process is one of reasoning back from the output desired to survey design.

14. This approach was a natural consequence of the context of data scarcity in which it was developed. Primary data collection is a slow and expensive process, so it made perfect sense to collect only data that could be guaranteed to be fully used for the production of statistics. This has led to a refusal to trust data that were not produced with a statistical goal in mind.

15. Today data are part of the landscape. They are on the Internet; they are produced by cars driving over detection loops, by mobile phones negotiating communication lines, by GPS systems signalling their position, etc. There will soon be chips recording the history of every item we buy, use and discard. It has become reasonable to ask whether input data should still be selected and edited to suit a predefined goal, or whether available data should participate in determining which statistical products will be relevant. We might, for instance, decide that it is preferable to produce international data straight away – in the same way as the GPI [3] – instead of producing national data first, and integrating them as a second step. This approach would have been awkward in the previous century, but the nature of data available today makes this perfectly reasonable.

16. The demands of modern society are completely different from what they used to be 20 years ago, and they are changing fast. If we make our output dependent on the availability of data rather than on predefined purposes, we might end up with a statistical process that is much more flexible. It will be able to react quickly to changes in society, because these changes will be automatically reflected in the input of our process.

C. Data editing

17. Primary data collection nearly always involves a manual process in which someone notes the requested data on a paper form or in the fields of an electronic questionnaire. Because this process has no added value for the data provider, there is no quality feedback, and errors inevitably creep in. It has always been recognised that this process needed error detection and correction. Detecting and correcting, however, involves high costs in terms of time and human work, and can be the source of new errors.

18. This has been one of the main points of focus of statistical methodology in the past century, and the forum in which this paper is delivered bears witness to the fact that efficiency and reliability of data editing remains under scrutiny and is under continuous development and refinement.

19. Data editing has become an art in itself. For many employees involved in it, it has become synonym for quality. Some of them perceive every effort to reduce editing activities as a menace for quality. So it can be anticipated that a proposal to rethink the relation between data editing and quality will meet resistance.

20. In confronting data editing inherited from the 20th century with the needs of the 21st century we need to address two questions:

- Do we need to edit all data?
- Is it appropriate to apply this method to large data sets?

21. The first question can be seen as an assessment of the quality of ubiquitous data. Data we access in secondary collection have their own goal. They are produced by normal social or economic processes. Errors can have disturbing effects on these processes, so significant errors may be expected to be detected and corrected within those processes. Therefore these data require much less editing effort on our part than primary data. We might even take this as a quality parameter of input data, and adopt a policy stipulating a preference for input data not requiring editing.

22. Statistics free of data editing already exist. The GPI [3], already mentioned in paragraph 6, is produced without any detection or correction of errors. An experiment by MIT economists shows the same edit-free approach [4]: they used Google Trends to predict fluctuations in house sales. They produced their data two months ahead of the National Association of Realtors, but their predictions turned out to be far more accurate.

23. The second question of paragraph 20 challenges the adequacy for large datasets of methods developed for primarily collected data. Modern data editing techniques focus on corrections that have a significant impact on the quality of the final statistical product. It is, however, a widely accepted premise that there is a direct relation between the quality of individual records and the quality of aggregates. Macro detection is normally followed by micro correction. This idea still lives, although we know in other domains that this relation does not hold. Photo editing programs, for instance, have a noise reduction function. Although applying this function produces a photo of better quality, the pixels of the corrected photo are no longer representative of the pixels of the original photo. But who cares about pixels? The same holds for audio editing programs: you can edit out scratches and coughs; this results in a better overall listening experience, but the original sound bits are lost.

24. We lack a theory of large data sets. Such a theory should enable us to distinguish noise from signal, or to identify other global parameters better suited to statistical data. It should enable us to determine quality in a way that is independent of the quality of individual records. It should enable us to create algorithms that enhance the quality of macro data – even at the cost of introducing distortions at the micro level. I am convinced that one day such a theory will exist. The question is only who will develop it, and when.

D. Trend continuity

25. It is natural to strive for trend continuity in a process that is purpose-oriented. If your work is determined by a predefined purpose, the output might as well have a stable relation to this purpose. In chapter III.B, however, we have argued that a process oriented on availability is more capable of responding fast to changes in user needs. In this context it is reasonable to reassess user needs, including trend continuity. How much importance do users attach to this need, and what is its relation to other needs? It might become acceptable that changes in the world leading to changes in our input are reflected in trend discontinuities.

E. Production by subject matter specialists

26. Human labour is a delay factor and comes at a high price. This is a reason for removing human intervention as much as possible from the production process. We can see that most industries have gone through such a transformation, concentrating human competence in a design process that creates an automated production process. If the output of this production process fails to meet quality standards, the process or the input is revised. The output is never corrected: it is either approved or discarded. Chapter V will show the feasibility of a statistical production process free of human intervention.

IV. Industrialisation: two examples

27. The last two centuries have been characterised by a profound transformation in the production of goods. Literature sometimes distinguishes two industrial revolutions. Some studies state that we are in the midst of the third industrial revolution powered by information technology. In fact, every type of goods has had its own revolution in its own time. On a global scale we have one great movement that affects all branches at different moments. Weaving was one of the first activities to be industrialised. The manual loom was replaced by a steam-driven machine. The weaving patterns were submitted on punch cards. So we see that from day one harnessing energy and information both played a part in the industrial revolution.

28. I would like to characterise the industrial revolution at the hand of two moments that I have witnessed through the emotions of participants that I knew personally: the furniture industry, which

changed dramatically in the fifties, and the Swiss watch industry, which was nearly swept away in the seventies.

A. The furniture industry

29. I was nine years old when I became aware that something dramatic was happening. My parents had just bought a table. They were very happy with it, because it was within their financial means, and they did not have to hire two strong men to bring it upstairs. Furniture in a box was a novelty then. They were sorting out the screws and figuring out how to put the parts together when my grandfather walked into the room and burst into a rage. He called my parents traitors. It was because of people like them that carpenters were going bankrupt. When the table had been mounted, his rage doubled. He would never have bought such ugly trash. He claimed that this unstable object would collapse under the weight of more than two people!

30. I have recalled his anecdote every time I saw something dramatic happen in the economic landscape, because it contains in a nutshell recurring elements of all industrial revolutions:

(a) A process of mass production appears on the market.
(b) The output of this process has an acceptable price. The time to market is short. The cost of ownership is low.
(c) A dramatic shift takes place in the quality criteria. Cost effectiveness and time to market come to dominate. Quality criteria that cannot be evaluated automatically within the production process are discarded and replaced by new, objective criteria. Users adopt the new criteria.
(d) Some traditional jobs disappear, new jobs are created. This presents both a threat and an opportunity.
(e) Some employees are not able to adapt. They lose their livelihoods and blame it on the users of the new product. They keep quoting quality criteria that most users no longer consider relevant.

31. It should be noticed that mass production requires an automated process. The employees involved in the new production process are no carpenters. Their skills involve monitoring the process. The subject matter specialists no longer produce individual articles. They only design them. Relevant competencies are removed from the production process and invested in the design process, which leads to a productivity boost.

B. The Swiss watch industry

32. Twenty years later I saw the same thing happen to the Swiss watch industry. The revolution in watch making is usually called the quartz revolution, but many Swiss call it the quartz crisis. Strangely enough, the Swiss pioneered the quartz watch, but they failed to see its economic potential, so the market was taken over by American and Japanese factories. Quartz spawned a mass production process, the output of which was affordable. The Swiss quality of mechanical watches gave way to the low cost of electronic watches from aboard. Thousands of jobs were lost. This, again, was blamed on the traitors who bought foreign trash instead of Swiss quality.

33. The Swiss watch industry was saved by the Swatch, which regained a significant portion of the market. However, the jobs of employees involved in putting mechanical movements together were lost forever.

C. Lessons for NSIs

34. These two examples illustrate what happens to a trade of hand crafted products when similar products are offered as the output of a mass production process. A question that leaps to mind is how could they be so blind? Didn’t they see it coming? Couldn’t they have done something to boost their productivity?

35. The answer to this question is paradoxical. They did see a need for a productivity boost. They worked hard on it, and they were very successful. They only failed to see that a productivity boost was not enough. There was a need for a paradigm shift, and it came from elsewhere. No one expected the

public to buy goods that failed to meet traditional quality criteria. But the quality criteria had changed, and both carpenters and watchmakers were overwhelmed by the success of products that they saw as trash.

38. Today's statistical process is also the outcome of changes aimed at boosting productivity. Secondary observation has taken over a significant portion of the input, allowing for automatic collection of data coming over the line. Error detection is performed at the macro level in order to limit micro correction to errors that have a significant impact on the results. Employees in charge of applying these principles, however, are like the carpenters and watchmakers of the past century: they take pride in the quality of their product and take offence at the suggestion that their quality criteria might no longer be the right ones. They see mass production as a threat for their jobs, but feel relatively safe because they do not believe that users of statistics will fall for data that fail to meet their standards of representativity.

37. Mass production, however, is not only a threat to traditionalists. It is also a chance to renew the art of statistics. It is a methodological challenge to develop new quality criteria suited both to the data at hand and to today's user needs. The statistical process in the near future will presumably be quite different from what it is now. What is to be feared is not that the jobs of people retiring soon will become obsolete – I find it much more distressing that young employees are acquiring right now skills that will soon be much less relevant. It is urgent to reform our processes in order to protect the coming generation from learning techniques threatened by obsolescence.

V. A new approach to statistics

38. This paper identifies a problem. It does not propose a solution, because thorough methodological research is required in order to define such a solution. The approach described in this chapter should not be seen as an answer, but as a hypothesis to guide research aiming at providing this answer.

A. Separating production from design

39. Mass production requires a process that is free of human contribution. This does not mean that human competence is not needed. Removing human competence from production means that all human competence is invested in design. We are not replacing humans by machines. We are building a relevant subset of human competence into the machine.

40. This, however, does require a redefinition of human competence. When a subject matter specialist is responsible for production, he or she can draw upon relevant knowledge and experience in every new situation. There is no need to anticipate on problems, because the person in charge is able to respond to a potentially infinite series of challenges. But when designing an automated process, it is necessary to anticipate on every possible situation, define the correct actions in advance, and identify any unanticipated outcome. This is a much more abstract activity, which is carried out in terms of processes and not in terms of products.

41. We saw the same skill shift two centuries ago in weaving: before mechanisation, a weaver needed to know how to weave a pattern into a material. Later on, he had to learn how to map a pattern onto a punched card.

42. So we should expect to see the emergence of two different categories of roles in the production of statistics: the producers and the designers. Producers know how to start a statistic process and to monitor it. They detect problems but are not responsible for correcting them (see the chapter on quality further on). Designers are in charge of defining the products and the processes that deliver them.

B. Separating process design from product design

43. One of the main problems with programs designed nowadays for supporting statistical processes is that they are not flexible. Changes in the product usually lead to changes in the process, which require a new version of the program. If we are to give support to mass production processes oriented on

available data, we need to devise a way of designing flexible processes, and to suppress all direct relationships between statistical products and IT products. This can only be done by separating process design from product design.

44. So we need to further split designers into two categories: Product designers and Process designers. Product designers design statistics. They refrain from designing the way these statistics are produced. They describe the relationship between input and output in terms of methods to be applied, and not in terms of steps to be performed. This approach frees them completely from any concepts or constraints from the IT domain. In their design they identify which aspects of the product are subject to change, so that the Process designer knows which parts of the process need to be flexible.

45. Process designer is the only role that has affinity with IT. A process designer is, however, not a programmer or software engineer. His responsibility is in the statistic domain. He has to design re-usable process steps and to combine them into processes capable of delivering the statistical products designed by his colleagues. He is the only statistical customer of the IT department. He orders IT products that support individual process steps. IT no longer has to produce systems to support a whole statistic process.

46. This approach leads to a significant reduction of the complexity of IT projects for statistics and greatly enhances their chances of success. It introduces a clear distinction between IT and methodological responsibilities in designing statistical processes. It puts an end to debates about whether a solution should be generic or specific, by confining specific aspects to product design and allowing process design to be carried out on a generic level.

C. Analysis and quality

47. Analysis is sometimes presented as an indispensable part of the statistical process. Without sound analysis, I am often told, there can be no quality. This is a serious argument, because analysis is pure human insight and cannot be automated. So if this is true, an industrial revolution in statistics is a phantom to be dismissed. I am, however, convinced that this revolution has already begun and cannot be stopped. So I will address this objection by describing the place of quality and analysis in a mass production process organised in the way presented in paragraphs A and B of this chapter.

C.1. Quality

48. In an automated process, quality needs to be defined formally and objectively. All subjective criteria need to be dropped. The first thing to do is to define quality criteria that can be measured within the production process. Product quality is defined in terms of user needs. Process quality is defined in terms of intermediary results and other process-oriented attributes. The proposed schema for the design and production of statistics allows for focus on quality at eight different points of the process:

(a) During product design: the product designer selects relevant quality criteria and specifies norms for them.
(b) During process design: the process designer uses an interactive simulation of the production environment and can play with test data. He does not ask whether he can produce good output, but whether he is designing a process that always produces good output. He delivers a process that computes values for quality criteria and confronts them with predefined norms.
(c) During acceptation by the product designer: he checks whether the process delivers the product that he has designed.
(d) During acceptation by the producer: he checks whether the process supports him in his monitoring task.
(e) If the production process does not start at the expected time: the producer inspects the process and the available input data in order to see which preconditions of the process are failing, and contacts colleagues competent to solve the problem.
(f) During production: quality indicators and their relation to their norms are part of the product, and are computed by the process. Abnormalities are detected and signalled. Good process design might even be able to build processes capable of taking corrective steps automatically.

(g) After production: the producer reads the quality reports delivered by the process and takes any measures required by the results. He never intervenes directly on the results. He might signal errors in the input data and request a new version of the input. He might request a revision of the process or product design. In all cases this will lead to a new run of the process. This may take place iteratively, until the quality indicators are in line with their norms.
(h) During analysis: yes, we still need an analysis step, but it comes at the end of the process and has no influence on the results. See next paragraph.

C.2. Analysis

49. Analysis has no direct influence on design or production. It only focuses on products and provides them with comments. It explains the how and the why of unexpected results. It makes the products accessible to a wide audience. The analyst may also detect undesirable properties of a product, and provide his colleagues with constructive criticism to improve both product and process design.

50. This role of analysis in a mass production process is clearly illustrated by the work of Google's chief economist, Hal Varian, in relation to the GPI[3]: the aim of his analysis is not produce a better GPI, but to help the public use it in a well informed way.

VI. The need for cultural change

51. This paper sketches significant changes that are taking place in the world. It has become clear that private companies can produce statistics that are similar to ours, but much cheaper and much faster. This will inevitably cause a shift in the quality criteria of users of statistical data. This situation requires a radical transformation in the way NSIs design and produce their statistics in order to keep up their relevance. This can only be achieved by a thorough change of culture.

52. The approach to statistics sketched in this paper is counterintuitive for statisticians used to design and produce statistics in the traditional way. It is contrary to the habits they have formed and to the assumptions that they have come to see as immutable facts. So it will require a extensive mind shift before they can embrace the new process. We have to pay attention to the following aspects:

- Create a sense of urgency: statisticians need to understand how the situation is changing and to accept that they have to adapt to changed paradigms.
- Reinvent quality criteria: NSIs need to accept that some quality criteria lose relevancy and to understand why this is inevitable.
- Build up knowledge and expertise in the field of industrial processing versus manual crafts.
- Innovation as a habit: NSIs need a change in working climate so that innovation can become one of their products.

53. Finally, NSIs need to collaborate on business concepts and architectures in order to achieve new solutions at affordable cost. The multiplier of international collaboration will help us to adapt to the changing world around us. There is a great opportunity to use the change process to converge on standards and concepts.

VII. Literature

[1] Kent, J.P: “Are we becoming dinosaurs?”, in Newsletter of the Sharing Advisory Board, May 2011 (Conference of European Statisticians, UN/ECE).
[2] Kent, J.P. & Wings, H.: *Official Statistics: Reasons for Change* [in Dutch], internal memo, Statistics Netherlands, February 2011.
[3] Harding, R: *Google to map inflation using web data*: http://www.ft.com/cms/s/2/deeb985e-d55f-11df-8e86-00144feabdc0.html.
[4] Popper, B.: *The New Leading Economic Indicators: Twitter, Google and Craigslist*: http://www.bnet.com/blog/high-tech/the-new-leading-economic-indicators-twitter-google-and-craigslist/345
[5] *Forecasting Financial Crises (FOC),* http://www.focproject.net/.

[6] *ICT research: Commission-backed project to help identify systemic financial market risks:* http://europa.eu/rapid/pressReleasesAction.do?reference=IP/10/1344&format=HTML&aged=0&language=en&guiLanguage=en.
[7] *London Investment Firm to Launch Twitter-Based Hedge Fund:* http://mashable.com/2010/12/16/twitter-hedge-fund-stock-market/
[8] Bollen, J. *et al*.: *Twitter mood predicts the stock market:* http://arxiv.org/PS_cache/arxiv/pdf/1010/1010.3003v1.pdf.