Industrialisation of Statistical Processes, Methods, and Technologies

Apted, Lisa; Carruthers, Philip; Lee, Geoff; Oehm, Daniel; Yu, Frank
Australian Bureau of Statistics
45 Benjamin Way
Belconnen ACT 2617, Australia
E-mail:lisa.apted@abs.gov.au;philip.carruthers@abs.gov.au;geoff.lee@abs.gov.au;
daniel.oehm@abs.gov.au; frank.yu@abs.gov.au;

During the industrial revolution so-called "cottage industry" approaches were replaced by standardised processes, new methods, and innovative technologies which had far reaching impacts on the efficiency and effectiveness of manufacturing production. The revolution created an environment in which outputs could be delivered more easily and cheaply. The outcome was not only a reduction in the resource base required but also an outpouring of new possibilities for utilising the products made possible by their increased availability at reduced cost. The "industrialisation" concept - of standardised production processes, supported by new methods, and made feasible by technological advances which allowed replication on a large scale - worked in a manufacturing context and there are sufficient similarities to suggest it should work equally well in a statistical production context. Here, it is assumed that cheaper access to greater quantities of high quality statistics will encourage the generation of more innovative presentation, reporting and analysis methodologies.

In the official statistical context, it's important to understand that industrialisation doesn't represent the robotic automation of standardised statistical activity. Rather, it offers improved capacity to replicate basic processes, freeing up analyst resources so they can add value where needed, with the support of knowledge-based decision-making controls where possible.

There are many different statistical processes within the ABS that are amenable to greater industrialisation, including the seasonal analysis of time series, microdata confidentialisation and the compilation of price indexes. Since at their core, all of the production processes for official statistics are about manipulating and quality assuring information, information management must be a strategic focus for greater achieving harmonisation and "industrialisation". Standardised metadata can be used to drive business processes and capture data relationships to promote greater harmony between similar or dependent activities.

ABS Examples

This paper explores three practical examples of "industrialisation of statistical production" in the ABS, from the methodological perspective. Notwithstanding the focus in this paper on the methodological perspective, it is important to recognise that new methods alone are insufficient. The desired outcomes can only be achieved from the interaction of new methods with significant process redesign, and with major and ongoing advances in harmonisation of information management and computing systems,

Three examples illustrate some of the types of methodological opportunities and challenges from

a strategy which focuses on greater industrialisation.

A Seasonal Analysis and Adjustment Example - Parallelisation

This example looks at the methodological opportunities which arise from changing a set of tasks currently done sequentially (seasonal re-analysis of a set of time series) to one where the component work is done in parallel (where the series are analysed as a block).

It looks at how an "industrialised" seasonal adjustment process in the ABS might require new methods which allow multiple series from a collection, or group of potentially related collections, to be examined simultaneously, based on statistical or data relationships defined in time series metadata or "series knowledge". Aligned processes will mean that activities, diagnostics and control variables will be centrally available so each process can influence the other, such as the detection of a large extreme observation in one time series will reinforce the evidence for a similar extreme value in a related time series. Industrialisation improves coherence and increases capacity and has the added benefit of creating the opportunity to free up analysts. That is, increased use of metadata to report quality diagnostics supported by the development of automation rules for the detection of problems will reduce the need for intervention by experienced analysts.

Time series in the ABS are currently adjusted using a knowledge-based metadata management system that observes a single time series at a time. Considerable amounts of metadata describing the data, analysis and outputs are collected and stored against each series but in a proprietary manner and in isolation of any other metadata repositories, including the major ABS information warehouses. The ABS proposes to adopt an aligned approach that departs from the current "series by series" method employed by the current ABS seasonal adjustment business process, where the expert analyst reviews a group of many series one at time, usually once per year. Aligned seasonal analysis takes a group of related series, along with their metadata, and analyses them, subject to layered sets of rules that facilitate automation of the seasonal adjustment process and associated quality checking. In one example, capturing the business rules for undertaking the initial pass over a time series would support automatic benchmarking of new time series. Any series discovered by the seasonal analysis system could be consumed and processed by the system and a set of preliminary diagnostics dispatched to the analyst and data owner automatically.

There is much to be gained analytically from the parallel seasonal adjustment of time series. Analysing a group of related series in parallel is not only faster, but the analyst need only concentrate on those series where comparison diagnostics indicate that attention is warranted. New metadata captured about the relationships between time series will be reinforced by new diagnostic metadata captured during analysis. The challenge for the ABS is to understand what metadata is needed for implementation and to make it work productively. While this is work in progress, early thoughts suggest that existing series knowledge will expand to include (amongst other things):

- Details of aggregation relationships, both temporal and contemporaneous;
- Relationships between time series estimates within and between collections;
- Control variables to delineate and control various steps of the seasonal adjustment process. For example, flags for problematic series, start analysis, finalise analysis.

Presently, metadata available during the analysis phase is insufficient to allow the adjustment process to make informed changes to parameter settings for one series based on another. Similarly, there is no capacity for metadata relating to other time series to be discovered and compared with that for the series being analysed. This potentially generates deficient outcomes, including mismatched

analysis settings, reduced coherence and a continued need for analyst intervention.

Industrialisation of seasonal analysis methodology is about business process improvement, with metadata from all sources to be considered and utilised. Pivotal to industrialisation is the increased use, maintenance and creation of metadata (series knowledge) that is also exchanged with other systems via recognised standards like SDMX (possibly extended to accommodate new metadata requirements about the details of the seasonal adjustment process). Increased integration with relevant data repositories and metadata registries will further enable the automation of the analysis and adjustment processes. Implementation of these new metadata standards and closer integration with ABS metadata registries/repositories will increase externalisation, accessibility and re-use of metadata. Access to time series will more closely aligned with pre-packaged ABS statistical products and macrodata table builders such as ABS Dot Stat. Users wanting to generate tables or new times series will be able to identify data sources and describe output requirements in terms of series metadata made available online.

Initiation of the aligned seasonal review process will be dictated by metadata control variables that are activated when the required estimates have been published. The ABS time series analyst will be alerted when a review of a collection of time series commences when original estimates and their metadata automatically load from repositories into the seasonal adjustment system, all subject to and controlled by ABS security protocols. Summary output diagnostics sent to analyst for review will flag those time series (if any) which are problematic. Information about the time series and its analysis will be sufficient to allow the expert analyst to determine whether to further investigate the any or all series. This enables the analyst to focus on more problematic or complex adjustments with less problematic series being finalised automatically. Finalised estimates of derived data would automatically be available to the ABS data and metadata dissemination platforms.

A Prices Indexes Compilation Example - Harmonisation of processing

This example explores how re-engineering the processing approach, and harmonising where possible across (currently) different production streams can enable existing methods to be applied in new settings. Industrialisation of price index processing will streamline index construction and publication.

The ABS is experienced in delivering high quality price indexes, but with a more structured approach gains could be made through efficient index compilation. For largely historical reasons, ABS has 4 different systems for producing each of 4 families of price indices. The four major ABS indexes: Consumer Price Index, Labour Price Index, Producer Price Index and House Price Index, presently use four different methods of calculation. Harmonising them and using a common information management framework will provide greater consistency in methodology and quality management.

At present collection of data for the price indices is operationally separate from other data collection work in ABS. Historically this made sense, but as automation steadily becomes more feasible for the collection of price data, the current processes appear increasingly inefficient.

Imputation, outlier management, quality adjustment and editing are procedures that account for a large proportion of the available resources and that could be described in metadata and automated. Replacing the manually intensive current practice of checking price observations to ensure they are correct and their movements are representative of the actual price movement, editing unexpected movements or confirming their accuracy with support from an automated system would vastly reduce

the workload associated costs of producing price indexes. The ABS is currently researching the feasibility of significance editing for the Consumer Price Index, Producer Price Index and House Price Index. Significance editing makes an important contribution to pricing processes by determining which observations have the largest impact on the related index and deserve high priority treatment. Optimised cut-offs captured in procedural metadata can be used to ensure Prices staff only check the observations with the highest priority. This methodological editing enhancement maximises the pricing staff experts time and efficiency. Significance editing has been implemented in the Labour Price Index and has shown to reduce the overall cost of the procedure. Price index compilation is industrialised by the implementation of an improved management framework and associated software system to produce all of the four key indexes.

There are additional benefits to industrialising the compilation of pricing metadata in a central location. Currently, index compilation is isolated from the time series analysis process. Newly defined indexes, once authenticated, could alert the seasonal adjustment system to undertake checks for seasonality automatically. From within the seasonal analysis system, it would also be possible to consume the metadata defining and describing the elementary aggregates, allowing compilation metadata to inform time series aggregation metadata. Analysts would also have a ready source of metadata to support drilling down from aggregates to components. The harmonisation of these two systems would permit an analyst to operate on low level data that they would not normally be allowed to observe, operating only with metadata-driven but trusted processes.

A Confidentialisation Example – User customization of statistical table production

This example explores the methodological advances needed to support a fundamental shift in the process design, from one where the statistical producer makes all the decisions (and does all the work) to produce output tables and analyses, to one where the consumer can specify precisely what output they require from a collection, when and as they need it. This will provide much richer and more flexible access to the end consumers of statistics.

A typical household survey file might contain hundreds of data items/variables. Simple combinatorial calculations show that there are an enormous number of combinations of possible output tables (or other analyses). Producing and disseminating all potential tables is infeasible, and would in any event lead to a significant problem managing all the tables and information. In a microdata context, the ABS believes that putting the user in direct control ensures that they get a better product. But ABS legislation only permits microdata to be released in a manner not likely to enable identification. Confidentialising a complete microdata file in concept requires a review of all potential tables, to identify which ones are "sparse" in ways which might expose individual records to identification. In practice pragmatic solutions have been employed, but these are time consuming, costly, and unsatisfying from a methodological perspective.

The ABS has invested in the development of an on-site tool called Remote Execution Environment for Microdata (REEM) to replace the existing intensely manual process of assessing and confidentialising microdata files. Rather than confidentialising the microdata file, the end user is able to specify, from the many potential tabulations (or analyses) which potentially exist, the outputs that they actually require. After this step the output is confidentialised automatically, in ways which can be replicated (so that the confidentialisation cannot be broken by repeated table requests). The confidentialisation method has been described elsewhere (Fraser, B. and Wooton, J. 2005).

The "industrialisation" innovation here is to implement that method using technology (a software tool) that can be applied to any dataset which has been described in a standard way. Providing

access for a researcher to another micro dataset requires only that the underlying micro dataset has been described properly, after which the user can select the tabulations of interest and so forth.

The key features of the final REEM architecture will be clear separation of the statistical process from the structured metadata which describes the data on which the process acts. This will be facilitated by the use of a central metadata registry and associated repositories. Importantly, metadata will describe the files on which the user is operating, and will record the shape of the output that the user receives. The automated confidentialisation method implemented in this way will produce output tables without further scrutiny by ABS analysts. This empowers users to consume greater amounts of data for their own real-time analyses.

The new functionality provided by REEM also allows for better monitoring of what is happening to data directly. For example, issues such as security and user table building preferences can be addressed.

Some Questions

How big is the Official Statistical Industry?

The examples above implicitly suggest that the "industrialisation" process (harmonising processes, methods and technologies) can occur within ABS alone and in isolation. This is misleading. Official statistical agencies have harmonised statistical frameworks (such as the System of National Accounts) across the world for many years. Moreover, there are strong similarities in the processes used by official statistical agencies across the world. The Generic Statistical Business Process Model articulates them at a high level. Finally, many official statistical agencies across the world face strikingly similar pressures and challenges as well - see for example Pink, Borowik and Lee (2010).

Presuming we can further standardise our processes well enough, can we move towards industrialising our methods and the supporting technologies as well? There are sufficient examples in this internet age where technological solutions have been standardised across an industry (e.g. air travel) to know that aspect is possible. Technological standardisation has occurred even when companies within the industry are competitors. The challenge for methodologists around the world to consider is whether and how we can "industrialise" our methodologies more. We share ideas well, at workshops and conferences such as this one - how can work together to make joint decisions about the best approaches for our "industry" to follow? And what are the most workable mechanisms for including the process designers and the technologists, not just in individual organisations, but around the world, in the discussions about harmonisation of methods that we do hold?

Must standardisation curtail innovation?

Anyone who has been involved in the process of developing an international statistical framework will be aware that can be a slow and tedious process. Reducing the rate of innovation in the development of the methods used to produce official statistics is a genuine risk for the "industrialisation" philosophy. It need not necessarily be so, as the experience of the growth of the internet has shown. The internet has created an environment within which enormous bursts of technological creativity and innovation occur - and yet at its heart it is reliant upon some basic standards which have been widely adopted.

The task for methodologists working in in the production of (official) statistics therefore becomes one of establishing which methods can and should be standardised, and which should not. Hall and Johnson (2009) propose a framework based on two dimensions to decide where artistic

(specially tailored or unique) processes make sense and where standardised processes should be applied. They also address the issue of how to manage artistic processes alongside standardised ones. Their logic may be applicable to industrialising methods as well.

The two dimensions are (a) process environment and (b) value of output variation to customers. Process environment considers the degree to which inputs and environmental issues are not uniform and hence require a craftsperson's adjustments. Value of output tailoring explore whether customers value distinctive or unique output, i.e. whether or not the customers want homogeneous products and services. Hall and Johnson recommend an approach which is summarised in table 1:

Table 1

	Process Environment	
Value of output variation to customers	*Low variability*	*High variability*
positive	Mass customisation	Artistic processes
negative	Mass processes	Broken processes

For mass processes, artistic discretion should be eliminated. This represents a situation where the customer desires uniform, homogeneous products and services, and the process environment would support their automated production - hence highly standardised methods would be valuable. The time series example does not as yet fit into this category (there is still too much variability in the environment) but it is approaching this state.

For mass customisation (producing controlled variations in outputs from uniform inputs) the outputs are limited to combinations of predefined components, and ideally individual customers have control over the particular variation the desire. This represents the best of both worlds of control and variation. The confidentialisation example presented earlier is following this strategy, and again, standardising methods are central to its success.

Broken processes are labelled as such, since the customer desires low variability, yet the input environment delivers such high variability that the output is too variable (usually despite the best efforts of the methodologists trying to reduce the variability in the process outputs). In the routine production of (official) statistics, this represents an undesirable and inefficient situation. One logical response would be to seek to reduce the variability in the inputs, perhaps using the philosophies pioneered by Deming and others, and then standardise the methods used to process those inputs, ie to migrate towards a mass process situation. The regular production of price indices might be categorised as a broken process which is following this prescription.

For artistic processes, which should only be chosen in the certainty that customers do value the variations, the direction is to leverage variability in the environment to create variations of products. An alternative response to broken processes might be to convince customers to value output variability to the extent that they are willing to pay for the effort involved in using the artistic process approach.

For official statistics customers generally are unwilling to pay on an ongoing basis, for artistic processes, and certainly should not be required to pay for broken processes. This suggests that the end goal the producers of official statistics should be to employ mass processes for routine outputs, supplemented with mass customisation where users themselves specify their individual requirements

as they dig deeper into the available information. Both require standardised, repeatable methods which can be automated.

Broken (so called) and artistic processes do however have an important role to play. They represent the stage in the evolution of a process or method where innovation can occur. For new statistical requirements, customers have shown themselves willing to pay the premium price involved in developing novel statistical products and services that they need urgently. For methodologists working in the official statistical industry, this is the stage when innovation can and should occur. The challenge is to recognise that (perhaps unlike academia where creativity, scholarship and invention are ends in themselves) for official statisticians this is an evolutionary step which will eventually be followed by a need to harmonise and standardise the new methods, so that the mass process or mass customisation route can be followed. Innovation is itself a process, which begins with invention and creativity, but should move towards a standardised, harmonised, repeatable, automated (ie "industrialised") end goal.

Summary

There are benefits from "industrialisation" - harmonisation of processes, standardisation of methods, and development of new ones which can be run in parallel and implemented in automated processes which can be replicated and run in parallel. Due to the automation, skilled experts will have more time to spend on data requiring the greatest scrutiny. Greater amounts of data can potentially be produced and consumed by users. Additionally, relevant diagnostic, quality and process metadata can be stored and managed centrally and consistently and become available for sharing with other processes. This will inevitably lead to greater opportunities for consistency of processes and coherence of data. Improvements in efficiency and quality gains by harmonising the processing as much as possible, creating single systems which can accommodate current and future data requirements such as seasonally adjusted data, trends or price indexes. Current processes will be integrated and standardised to ensure common terminology, common imputation methods are used for missing observations, common graphics and other outputs meaning a clearer understanding for users.

REFERENCES

Joseph M Hall and M. Eric Johnson (2009) (“When should a process be art, not science”, published in Harvard Business Review, March 2009 issue, pp. 58 – 64)

Fraser, B. and Wooton, J. (2005) (“A proposed method for confidentialising tabular output to protect against differencing”. Presented at UNECE Work Session on Statistical Data Confidentiality, November, 2005)

Pink, Borowik and Lee (2010) (“The case for an international statistical innovation program – Transforming national and international statistics systems” Published in the Statistical Journal of the International Association for Official Statistics (IAOS). Volume 26, Number 3-4 / 2009/2010)