

National Statistics Offices and the Prosumer Challenge

John Ellenberger¹, Stuart Muir²

¹Space-Time Research Pty Ltd, 1/601 Church St Richmond, Victoria 3122, Australia.

e-mail: john.ellenberger@spacetime.com.au

²Symbolix Pty Ltd, 1/14 Akuna Dve, Williamstown North, Victoria, 3016, AUS

e-mail: smuir@symbolix.com.au

Abstract

National Statistics Offices (NSOs) are under increasing demand to provide greater access to official statistics for their key stakeholders and the public at large. Timely, relevant, and robust statistics are recognised as fundamental economic and social measures of an economy's performance.

Historically, NSOs have done well in servicing public users with aggregated statistics: however the needs of more demanding users, researchers, policy makers and analysts tend to be met through clumsy data laboratories or expensive custom table production systems.

This group of "Prosumers", while being small in number, tends to be the greatest user of statistics and the ones who are faced with lack of access to detailed statistics. Hans Rosling from the Gapminder Foundation is considered a quintessential Prosumer.

This paper focuses on the rise of Prosumers and the ways in which providers of statistics are beginning to service their needs. We are beginning to see a greater recognition of the role of the Prosumer, who takes statistics and turns them into interesting stories for policy makers.

We will consider the key challenges and opportunities that NSOs face in servicing this evolving market. New confidentiality routines, new Internet based applications allowing users to query Micro-data, faster tabulation engines to cope with large databases and above all an increasing recognition of the role Prosumers play in the statistical value chain removes the barriers associated with servicing Prosumers while still ensuring that the privacy of respondents and integrity of the data is preserved.

Keywords: Micro-data, Confidentiality, Dissemination

1. Historical background

Note: This paper refers to NSOs, but because there can be many statistical producers in a country; the reference to NSOs is to all organisations that produce official statistics.

Traditionally the core role of the NSO was to support policy generation, facilitating this through transparent and robust statistical analysis and politically unbiased

interpretation of the same. Although this role has evolved over time, NSOs continue to:

- facilitate broad and deep information access
- protect that same information from misinterpretation
- protect the privacy of the individual data source.

Up until twenty years ago, the role of a central statistical organisation was dominated by the major needs of the Government it served. The delivery of statistical data then was rather rigid, and heavily production orientated.

If we view a statistical data service of those times as having three dimensions of **timeliness**, **robustness**, and **coverage** as in Figure 1), then typically an end user was constrained to a choice of any two. A timely and robust table would necessarily lack depth, whereas a deep and robust analysis would not be timely in its delivery.

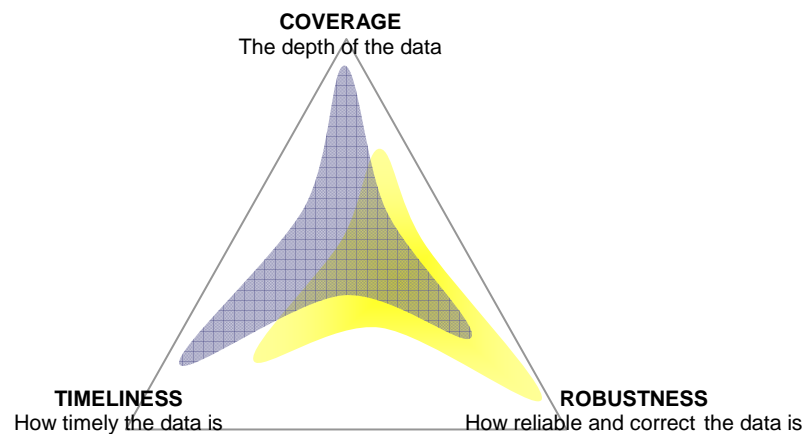


Figure 1: Visualisation of the dimensions of the NSO's historical data provision.

This service model evolved partly because of the data consumption market place, and partly due to the philosophical and technological environment of the National Statistical Offices.

1.1. Traditional user groups

Traditionally, NSOs dealt with two key groups: *consumers*, that is, the general public; and *professionals*, particularly government analysts and statisticians. An implicit duty of care was extended by the data-controlling agency to its clients. This was discharged differently for each of the two core groups.

The general consumer was the larger group, with needs that could be satisfied with aggregated data. This data was generally supplied through standard (sometimes hard copy) tables of crossed variables. Such a system allowed the NSO to adequately control issues such as confidentiality and applicability through simple control of the published tables. A typical census report, for example, might consist of a few hundred tables.

The meta-data that ensure comparability of data could be clearly attached and connected to each table, allowing the NSO to discharge its duty of accurate interpretation of the data.

Such tables could be generated in a timely fashion, and be assured of robustness, yet could not be particularly deep. Fundamentally, these standard tables suffered from problems of applicability. If a requirement wasn't identified early on in the production process, then the NSO needed to initiate a customized and manual extraction process to generate the requisite information. This effort had to be repeated whenever different data requests were made upon the NSO.

This shortcoming made standard tables generally inadequate for the second core group: the researchers and analysts. These professional data consumers were more interested in access to the micro-data. They needed to perform analysis on particular subsets and groups within the larger data set.

In theory, at least, the professional analysts were versed in statistical methods enough for them to carry some onus of interpretation. The NSO was pressured to allow unfettered access to the micro-data needed by this particular power user group to perform their duties.

Here, the NSO faced new challenges in protecting the identity of the individual data elements, whilst allowing the necessary analysis and research to continue. Typically, this would be achieved through a 'licensing' system, where stiff penalties were imposed for a breach of conditions.

There is a lot of material available for the interested party to study regarding the issue of Statistical Disclosure Control, (for example Willenborg and Waal, 1996), but to use the apt terminology of Buzzogoli and Biggeri (2001), the two core principles reduce to either 'safe data' or 'safe setting'.

With 'safe data' the data released is itself controlled, through the use of aggregation, suppression or obfuscation methods to protect the individual identity. We also extend this concept of safe data to include the requirement that all the required meta-information needed for its correct deployment and interpretation is adequately attached. The 'safe setting' model involves control of access to the data, not the control of the data itself.

As the role of the NSO has evolved, the realistic adherence to such fundamental precepts becomes less easy.

2. The changing face of the NSO market

Advances in social sciences in particular, combined with the advent of the 'information superhighway' and its associated increase in the demand for information, have resulted in challenges to the control of national data by NSOs, the traditional custodians.

This demand for national data, for everything from comparison benchmarking to public policy evaluation, has led to a unique conflict of privacy versus access.

Without adequate protection of the privacy of respondents, future data collection will be compromised. Without adequate access, the return on the significant investments made in the gathering of the data will not be realised.

The original service model, as we have portrayed it above, was fundamentally a binary classification. Once a general consumer requested more information, their classification was changed to a professional, and they were assumed to be representative of this class of user with its advanced knowledge and needs.

The blurring of the distinction between the two groups that such evolution has necessarily generated has created a new type of data consumer – one who consumes data at a high level, but does so to produce more information at the lower levels, the *'prosumer'*.

Figure 2 shows these relationships.

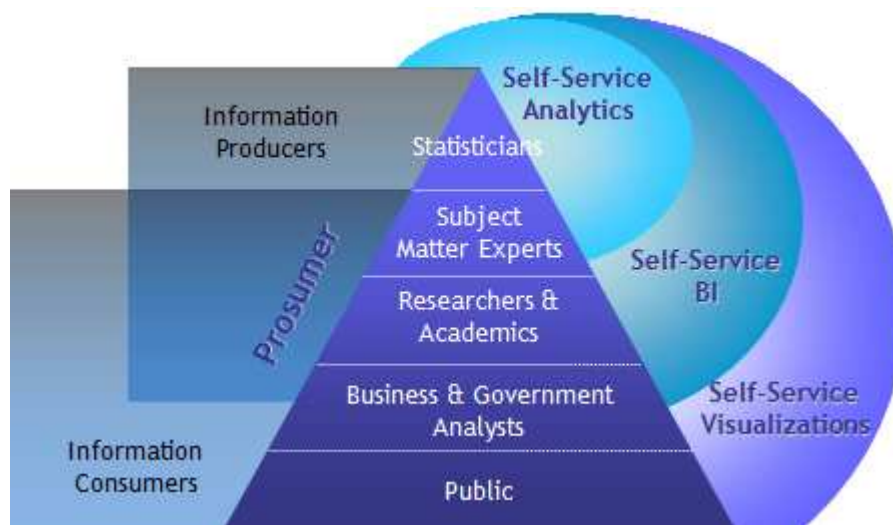


Figure 2: Market sectors of the modern NSO.

With the advent of new information technologies, the previously small group of academics and researchers who are typically the most demanding group on an NSO's resources now insists on faster and broader access to data. This group has evolved to what the Australian Bureau of Statistics refers to as 'harvesters' (ABS, 2009); regular data farmers who specifically target established data sets to support decision making, or to support a current hypothesis. They do not browse through the products on offer, but have established needs and requirements.

Conversely, the original general consumer body, traditionally satisfied consuming the produced standard tables, now desires more resolution. This may be as simple as wanting to know how their local area compares with some larger measure, but nonetheless this is a demand for applicability and timeliness that stretches previous delivery models. The ABS refers to these users as 'tourists', as they do not have structured technical needs, but are looking for general information and are not likely to be willing to push too hard to find it.

The middle ground is occupied by the prosumer, one who will invest significant resources mining for the specific nugget they are seeking. His needs are as unstructured as are those of the tourist, but are just as highly specific as those of the harvester.

The impact of the prosumer on NSOs can be witnessed in Statistic Canada's survey of 22 NSOs (Roy, 2005), which found that the top five web user concerns raised (in order of importance) were:

1. Ease of site navigation
2. Effectiveness of search capabilities
3. Availability of regional level data
4. Documented methodologies
5. Effective on-line retrieval

These are not the needs of tourists (with a passing interest), or harvesters (with a predetermined and regular need). These are the needs of a prosumer.

3. Profile of a Prosumer

The middle ground of the data consumption pyramid is now owned by a new class of user. These users, while small in number, tend to be the greatest users of data and the users most threatened by lack of access to timely detailed statistics. By virtue of NSO research, they have been classified into a user persona. For example the Australian Bureau of Statistics, calls them 'miners' (ABS, 2009). We adopt the alternative term 'Prosumers' (being both a professional supplier of information, and a consumer of data).

Here we see a new breed of analysts and statisticians from private enterprise, market research and journalism among others, working on high-level, specific projects. They maintain the vocal demands of the traditional analyst, yet do not necessarily have the technical training of this progenitor group. Furthermore, these prosumers may not be as socially motivated as the previous researchers. Their use of high-powered statistical engines is facilitated by the advent of easy-to-use software. Also, their incentive to find a nugget of information may be motivated by more commercial reasons. They are likely to be running their own analyses on data that is downloaded from the NSO. These analyses may need to go beyond publications and statistics from publically available data cubes, spreadsheets or time series. They might need access to record level data.

This increasing demand for access at the micro-level poses serious issues for all NSOs, who are increasingly required to grant free and unfettered access to data for unspecified applications, yet must still protect the same users from misinterpretation of the information, and from disclosing detail properly considered confidential.

At initial glance, the prosumer may appear to be a monster, challenging the principles, structure, and processes used to serve each previously neatly segmented data consumption environment, and bringing no benefits. However, the prosumer serves a facilitation role, consuming large amounts of data to produce accessible and relevant information for the consumer audiences, be they policy managers or just an

information-hungry general public. As such, finding effective ways to interact with and serve this market provides an incredible opportunity for NSOs.

3.1. Case Study – Gapminder

To see the incredible interest a prosumer can generate, and the benefits that such interest can promulgate, one exemplifies Hans Rosling and Gapminder (www.gapminder.org).

A prolific consumer of nationalised data, the technology of Gapminder makes the complexity of metrics approachable and, more importantly, visible. Such a tool was likely never intended to be employed by a lay audience. Its data requirements are too severe, yet its outcomes are precisely what its audience desires.

The compelling nature of information trends viewed through Gapminder can be seen directly through Hans Rosling’s presentation at the TED (Technology, Entertainment Design) 2006 conference. This presentation is ranked in the top 10 talks on the TED website (www.ted.com), and is the subject of much online discussion and acclaim. It has a 5 star rating on YouTube, a remarkable accomplishment for a presentation on statistics.

As well as providing accessible information through popular media, Rosling’s software is used in dissemination of official statistics such as the OECD Factbook 2008 (OECD, 2008).

Gapminder’s mission statement exemplifies the important role into which prosumers are positioning themselves:

“To promote sustainable global development and achievement of the United Nations Millennium Development Goals by increased use and understanding of statistics and other information about social, economic and environmental development at local, national and global levels.”

This philosophy lacks the focussed hypothesis testing of the professional harvester, yet desires data at just the same depth.

4. Towards a new model for data provision

As the tools available become more sophisticated, and the desire for such presentations grows, the NSO’s duty of care model (‘safe data, safe setting’) must evolve to both:

- protect the prosumer from erroneously ‘data-dredging’ a nonsensical finding
- maintain confidentiality of the source.

This competing demand of access to information at the micro-data level is the management challenge for the NSO.

The easiest way to understand this challenge is by exploring the prosumer user group and how its particular needs can be met through the application of specific new technologies.

4.1. The prosumer requires access to micro-level data

As it is rare for prosumers to commence research with adequate detailed specifications, and because they will be employing a variety of high powered analyses, the prosumers will demand access to micro-level data.

Due to their sheer numbers, it is not feasible that prosumers will sign strict confidentiality agreements, or agree to operate from within a physical safe room as was previously employed.

The solution here will be on-the-fly data obfuscation to protect individual data identities. This will allow the proliferation of Remote Access Data Laboratories, where the NSO maintains the physical data repository, and the user accesses remotely to retrieve vetted reports.

There are two main data protection techniques that we believe will ultimately be employed by the data provider, and only one will likely survive into future applications. These are:

- Suppression
- Modification/obfuscation.

Suppression is a favoured tool of many NSOs as it is a visible display of them enacting their duty of care. It is recommended by the ONS (ONS, 2006) for use with sensitive statistics (such as low birth weight children) and is in the National Statistical Service handbook (NSS, 2008).

However, suppression can hinder further analysis by physically removing not only the cell deemed sensitive, but also those that would allow its deduction. Additionally, the data quality characteristics of suppressed tables are generally considered to be poor (Cox & Kelly, 2005).

If we consider that many prosumers might be working on the mining of low count information, such as in disadvantaged neighbourhoods, then suppression methods will at the least need to be supported by other technologies, such as Random Rounding.

An example of such technology is Tau Argus. Tau-Argus uses threshold, dominance and probability rules to identify cells requiring disclosure control. The Tau Argus software can then apply control methods such as recoding, consequential suppression, and controlled rounding. The latter has the advantages over simple rounding of being more difficult to unpick, and producing totals that are additive. If suppression is chosen, several rules may be used at the same time to decide which cells require initial or primary suppression.

Because marginal totals are generally published along with the cell values, it is necessary to suppress further cells, called secondary cells, so that the original cell values cannot be calculated by subtraction. The person applying the disclosure control can choose from several methods for the suppression of cells which minimise information loss. The software package can be used stand-alone, or in conjunction with other software such as SuperCROSS (Staggemeier, Lowthian, & Lee, 2007)

Modification algorithms deduce sensitive cells and modify their value appropriately. The definition of sensitive data is not for this discussion, however once a cell is deemed sensitive, its value is perturbed sufficiently for conformance to confidentiality standards. There may be further perturbations through the table to preserve margin totals, for instance. An example of this is Controlled Tabular Adjustment (CTA) (Dandekar & Cox, 2002). Further advances to this technology mean that certain higher statistical moments can be maintained despite the perturbation (Cox & Kelly (2003)). This is an important feature for prosumers.

These measures amount to a technological manifestation of the “safe setting”, where access to raw data is allowed, yet its visualisation is controlled. This can be achieved through the web with an NSO maintaining custodial control over the physical servers and data-cubes, with the prosumer granted access through the “safe room” of the thin layer software.

4.2. The prosumer may not be a data expert

Whereas the academics and researchers of the previous era were trained in the abstractions and methods of statistical analysis, the modern prosumer may be relying on “intelligent” software. The duty of care that was previously discharged by an agreement, or the printing of relevant meta-information upon tables, can no longer be maintained.

Therefore the NSO needs a complete meta-information repository that is logically connected to the records to preserve the value of the data (Richter & Cornish, 1996). Information such as the reliability of the data or particular survey techniques can then be directly attached to any view of the dataset. Given the freedoms that are demanded (even if only demanded through economic cost of not providing them) this repository requires a complex connectivity to adequately protect an end-user from inadvertently relying on unsound data. For instance, many tables in New Zealand’s 2001 Disability and Services census were deemed by NZ Statistics to be unreliable, and consequently suppressed in the published reports.

4.3. The prosumers are numerous, diverse, and unruly:

Prosumers can be found in every walk of life, both corporate and public. As such they will be demanding access to previously unthought-of hybrid metrics and recoding. What will be applauded as a very helpful attribute by one group will likely be castigated as an unforgivable corruption by another group accessing precisely the same information for an alternate purpose.

To combat this, the technology must be able to deliver relative freedom to recode and define variables in the reporting layer. This has profound implications for the meta-information repositories, and the confidentiality protections.

Coded solutions that require intimate knowledge of the underlying data-cube are too slow for a prosumer. The large variety of questions, and specific queries that a prosumer might desire, whether to feed their curiosity or to supply a specific form and need (such as Gapminder input) precludes slow preparation and access. The opportunity here is to provide simple interfaces to extract the data readily, and to readily conform that view to any number of interests. For example, since 2000 all ABS publications have been made available electronically (Tam & Kraayenbrink,

2006) and this offering has expanded for the 2006 census offering to include a thematic mapping service (CData; <http://www.abs.gov.au/cdata>)

4.4. Prosumers are Vocal and demanding

Prosumers have grown from some of the more pressured areas of the corporate sector, such as commerce. Just like their predecessors, their demands stress the access methods to data. They are also pressured to find answers quickly.

A common tool of the data miner is the “question-answer-question-answer” style of ad-hoc mining. In this paradigm, the prosumer will actively slice and dice through the data-cube, following hunches hinted at in the results of their last query. This method can be fraught with the risk of so-called “data-dredging”, where one finds chance occurrences of patterns that appear quite statistically strong.

Although the NSO is unable to protect the prosumer from all possible manifestations of this, the meta-information will need to be adequately attached (again) to at least give the NSO some recourse and defence, given they will have no control over the actual queries now placed upon the data.

Without this freedom to slice in an ad-hoc fashion, the prosumer will rapidly deplete the already diminished NSO budget with their demands. The only recourse for the NSO is to push this demand on resources back onto the prosumer. This is best done through the application of light-weight, self-service style front ends.

Using a web browser as the interface allows a reduction in support costs, as web standards are employed rather than custom or unique software packages. In this way, more users can access the same data sets, complying with the “write once, publish many times” philosophy that drives returns on the collection of the set.

4.5. High-speed access and turnaround times

The prosumer is a creation of the information age. Whether operating in policy, or the private sector, they find timeliness is a key constraint. Parliament question times can raise issues rapidly, partially in response to public opinion polls, and the modern analyst will be forced to generate a robust evaluation of a policy in a short time.

Marketers also will be working off shorter and shorter turnarounds, demanding commensurate access times. As such, the older safe data style of building and selling data-cubes will become less viable. The internet is the accepted medium, and prosumers will measure return on investment less and less through the robustness of the data, and more through its timeliness and depth.

High speed access can be supplied through modern, columnar based repositories which can dramatically reduce query-times on large volumes of data. This access time of complex queries is cited in The BI Survey (Pendse 2009) for the fifth consecutive time as a leading concern of end-users. The evolution of web-based access to such high speed access engines is a welcomed technology by the marketplace.

5. Conclusions

The advent of information bandwidth, and advancement of social science techniques have combined to generate a redefinition of the core audience of the National Statistical Office, whilst demanding the role as custodian of data and protector of privacy is retained. The modern NSO is now expected to grant access, to an increasingly diverse group, to the large investment in data made by Governments around the world.

The evolution of an intermediate ‘prosumer’ of data products, will place unique demands on traditional service delivery models, but simultaneously provides unique opportunities for proactive data providers.

New technologies are vital to fill the current gaps between demand for access and information, and delivery of service. Key current technologies include:

- Suppression and modification algorithms to protect confidentiality
- Ability to control and disseminate large amounts of meta-information
- Relative freedom to recode and define variables in the reporting layer
- Thin layer software as a modern alternative to the “safe room approach”
- Light-weight, self-service style front ends to the data
- High-speed access and turnaround times to applicable data queries

By implementing appropriate software solutions, the NSO can satisfy these needs of the prosumer, protect an individual’s identity through on-the-fly statistical disclosure controls, whilst still allowing the analytical access to large cubes that justifies the expense of collating them.

References

- ABS (Australian Bureau of Statistics) (2009)
- Buzzigoli, L., Biggeri, L. (2001) Statistical disclosure control and data access for research purposes: critical issues and possible solutions, *Bulletin of the International Statistical Institute, 53rd Session, Proceedings*, TOME LIX, Book 1, 257-260
- Cox, Lawrence H., (2005) Quality-preserving controlled tabular adjustment: a method for resolving confidentiality and data quality issues for tabular data, *Statistics Canada International Symposium Series Proceedings*.
- Cox, L. & Kelly, J. (2003) Balancing quality and confidentiality for tabular data. In *Eurostat work session on statistical data and confidentiality*, Luxembourg, 2003 Available from:
www.unece.org/stats/documents/2003/04/confidentiality/wp.4.s.e.pdf
- Dandekar, R. & Cox, L. (2002) Synthetic Tabular Data – An alternative to Complementary Cell Suppression, *Manuscript*
- Fraser, B. and Wooton, J. (2006) A proposed method for confidentialising tabular output to protect against differencing, *Internal report*, Data Access and Confidentiality Methodology Unit, Australian Bureau of Statistics.
- Longhurst, J., Tromans, N., Young, C., Miller, C. (2007) Statistical Disclosure Control for the 2011 UK Census, *UNECE Work session on statistical data confidentiality*

- OECD Factbook 2008 in Gapminder Graphs
http://www.oecd.org/document/1/0,3343,en_2649_33715_40680833_1_1_1_1,00.html
- ONS (Office for National Statistics), 2006, 2006 Review of the dissemination of health statistics: confidentiality guidance, Office for National Statistics, Available from: www.statistics.gov.uk/about/Consultations/disclosure.asp
- NSS (National Statistical Services), 2008, *Handbook*, Available from: www.NSS.gov.au/nss/home.nsf/pages/NSS+Handbook?OpenDocument , Accessed 22/4/2008
- Pendse, N. 2009 The BI Survey [online], Available from: www.bi-survey.com [Accessed 16/1/09]
- Richter, W., Cornish, J. (1996) Meta data systems to turn numbers into information, *Seminar on Official Statistics – Past and Future, Session 5: Our legacy to future generations, Conference of European Statisticians*, Lisbon, Portugal.
- Roy, D. (2005) Consultation with National Statistics Offices, *Web Presence Strategic Vision, International Marketing and output database Conference*, The Hague.
- Andrea Toniolo Staggemeier, A. L., Lowthian, P., Lee, G. (2007) Applying Tau-Argus software to SuperCROSS Tables: a practical example using the UK Business Register Unit data, *Joint UNECE/Eurostat work session on statistical data confidentiality*, Manchester
- Tam, S. & Kraayenbrink, R. (2006) Data Communication - Emerging International trends and Practices of the Australian Bureau of Statistics, Australian Bureau of Statistics, Australia.
- Willenborg L.de Waal, T. (1996) Statistical disclosure control in practice, in *Lecture Notes in Statistics*, 111, Springer-Verlag, New York