IN SEARCH OF DATA INTEGRATION: NO MATCHES FOUND

Gordon E. Priest, Statistics Canada
R.H. Coats Building, Ottawa, ON, K1A 0T6

Key Words: Stove-pipes, Meta Information, Harmonization, Standards, Integration, Thematic

1. Introduction

The evolution of the statistical agency

Statistical activity began in North America both informally and formally. Informally by the Jesuits in New France and reported in the Jesuit Relations from 1632 to 1672 in which casual enumerations of the Aboriginal population were noted. The first formal census was conducted in the Commonwealth of Virginia in 1630 by the Board of Trade in London who no doubt wished to ascertain the robustness of the investments of its members. The Census of New France in 1655 was an example of a census conducted specifically to support demographic, social and economic planning. Herein lay the groundwork for the national censuses which were to follow the Declaration of Independence in the United States and the Durham Report in British North America.[1]

While censuses remained the keystone of statistical activity for many decades, the introduction of sampling methodology tended to expand the content of the census. Subsequently, the same methodology saw the development of sample surveys. Sampling was also implemented in the derivation of data from administrative records.

Many of the new statistical activities and vehicles grew around the needs of servicing the specific needs of specific clients. The very success of these activities, however, led to a fragmentation, even of highly centralized national statistical agencies. We developed censuses, post-censal surveys, household surveys, business surveys and administrative record data derivation. Thus, our data gathering evolved as a

---

[1]

The first census of the United States was conducted in 1790. While the first post-Confederation census of Canada was taken in 1871, the periodic colonial censuses of British North America were generally regularized and standardized commencing in 1851.

vehicle-driven system. And each vehicle tended to develop its own expertise in systems, methodology and subject-matter.

Islands, silos and stove pipes

The statistical agency became what Tapscott and Caston (*Paradigm Shift: The New Promise of Information Technology*, New York, McGraw-Hill, 1994) have referred to as "the problem of the unregulated enterprise." They describe islands of technology or expertise which meet specific needs but result in a fragmentation of the organization. They note that such islands have limited and specialized functions that may have nothing to do with overall business objectives or strategies of the corporation. Furthermore, the islands may become balkanised with formidable physical and organizational barriers, redundancies and inefficiencies. They cite lack of integration as a source of significant loss in business opportunities.

Keith Vozel of AT&T , in his "Technical Evolution White Paper", described such organizations as vertical or stove-pipe, the parts of which tend to address a single issue or client without regards to the needs or requirements of others. These organizations are wasteful in terms of redundant or replicated data in which there is no enterprise or corporate view of the holdings. Other literature refers to such organizations as silos to which access is difficult and between which communication is non-existent or limited. These silos represent untapped potential and lost opportunities.

Other critics have described the statistical agency as an organization of autonomous data development programs. My own view of statistical organizations is that they are less corporations than they are consortiums of independent producers. While many of these producers have well-served their specific clients, it has not been without a price.

2. Implications

Lack of meta information

Statistical agencies generally have very little, if any, *corporate* knowledge regarding the nature and extent of

their data holdings and what knowledge they do possess, has not been systematically shared with clients and potential clients. How often have we heard a policy maker, decision maker or researcher lamenting the lack of data when suitable data actually existed but were buried away from sight in some antiseptic and air conditioned tape library? Unfortunately, the production of meta information (that is, information about the data holdings), is very dependent upon the various production areas. The amount of meta information that is held may vary significantly from area to area and it is not usually documented to any corporate standard. Where attempts have been made to develop standardized meta information it is more likely to serve some bureaucratic purpose rather than potential clients. This results in under-utilization of the data collections. Clients, as well as agency staff, undertaking research on any given issue or population, are left largely to their own devices to contact *each* of the source areas to determine if any relevant data are available. The task is formidable, frustrating and often, fruitless.

Disharmonies

As might be expected, given the nature of independent production, further complications exist due to disharmonies between vehicles or sources in terms of concepts, definitions, classification systems, and documentation. Not only has each production area developed its own methodological, processing and dissemination practices, so has it developed its own subject-matter content. Through lack of care, communication or perhaps resources, differences have arisen in terms of concepts, definitions, classification systems and database coding. Not only is this distressing to the end user but it is also wasteful of resources. Given the lack of corporate standards, program managers, time and again, develop totally new documentation, unmindful of what might already have been produced elsewhere in the agency.

We are all no doubt aware of those situations where a data set from one source cannot be compared with another source, even though it bears the same name. On the other hand, there are those cases where variables actually are comparable but carry different names. At Statistics Canada we have even uncovered situations where variable names may be comparable in one official language but not in the other. And we have probably all experienced those situations where, even though a variable may carry its conceptual integrity from one source to another, comparability may be lost because

each source used a different classification system or used non-standardized aggregations. Finally, there are those insidious practices of using different mnemonics in the coding of variables on micro data file record layouts. This can lead to serious coding errors for persons working with multi-source files.

Contradictory or incomplete outputs

Another legacy of our stove-pipe production is that of independent vehicle-driven output. There are the obvious difficulties when Survey B contradicts the earlier released figures from Survey A. Such incidents are followed by the usual flurry of releases containing footnotes and qualifications explaining that one source was seasonally adjusted, or was rounded to prevent residual disclosure. Or sometimes we just issue a blushing pink errata sheet and *'fess up'* to a "computer error". While it is understandable that estimates from one source may not equate to estimates from another source, failure to document such differences is inexcusable.

Single-source output biased

Of greater concern is the analytical output that releases a set of information from a single source without the benefit of related and relevant data from other existing sources. Such releases can be dangerous in terms of providing partial and therefore, biased and misleading information. That is, the information is not set in the context of our comprehensive knowledge of a situation.

Implications of stove-pipe production

To summarize the implications of stove-pipe production in statistical agencies, we see that the corporation's knowledge of the extent and nature of its data holdings may be incomplete and therefore, of diminished use to the client. Disharmonies exist between sources and, therefore, even when the client does find different sources of interest, the data may not be comparable. Finally, the agency may mislead clients by releasing vehicle-driven data rather than integrated outputs.

If we accept that fragmented production poses a problem for clients then we have to consider integration as a solution. That is, we must start with a corporate inventory of our holdings (meta information), we need to resolve the disharmonies and we need to ensure that data releases are made in the context of our full knowledge of a situation.

Compelling reasons for action

There are compelling reasons to take these actions now. Firstly, many agencies are faced with funding cuts at a time when the demand for information is increasing. It is understandable that in tough economic times policy makers and decision makers in both the public and private sectors want the most reliable, most recent data because the implications of making a wrong or uninformed decision are far more serious. It falls, therefore, to the statistical agency to not only do more with less, but to work smarter and that includes mining and utilizing existing data as fully as possible. And you can't mine what you don't know you have. Maintaining dynamic corporate meta information and metadata just makes good business sense.

Secondly, technology now exists to make the job of data and metadata management infinitely easier than was the case ten, or even five, years ago. Hardware is faster and has greater capacity, networked computers make the sharing of information easier and software is much more user-friendly.

Thirdly, clients, especially those with Internet experience, have become increasingly knowledgeable and sophisticated with respect to searching for information. Thus they have increasing expectations of being able to approach a statistical agency, browse its holdings, specify output and download it; online, real-time at low cost or no cost. While there will be undoubted costs in building such a service capacity there is also a potential for hard cost reduction (cost avoidance) and improved productivity. For example, agencies should reduce the number of expensive generic products and allow, encourage or assist clients to build their own niche products.

3. Future actions

The vision

Thus, there is need and there is opportunity. We must develop the vision and the corporate will to accept the challenge and seize the opportunity. There are three fundamental components to the vision. Build the meta information and provide access to it, resolve the disharmonies and move from vehicle-driven outputs to issue (or population)-driven integrated outputs.

Building the meta information

Meta information must be comprehensive. It must respond equally to the client who simply wants an answer to a question such as the number of widgets produced last year as well as to the client who wants to know what is resident on micro data bases so he or she can do his or her own research. Therefore, meta information must describe the contents of micro data files, the contents of aggregated tabular output, the content of analytical or descriptive reports and the nature of specialized services provided by the agency. The information must be accessible by a search tool that facilitates both keyword and thematic searches. Ideally, a thesaurus should sit in front of such a tool to translate the client's lexicon to the agency's lexicon. The importance of a thematic search tool cannot be underestimated as is witnessed by many of the more helpful sites on the Web. The listing of subjects or themes and variables associated with those themes enhances the search by revealing variables that may be useful but not previously evident to the client. Regardless of whether the client searches on the basis of keyword or themes, however, the outcome should be the same. That is, he or she must be directed to the *source* of the information or data sought.

One gateway: one tool

Experience has shown that clients have found the statistical agency to be a bewildering maze of seemingly illogical sources. How many of us working in statistical agencies have had calls that were prefaced by, "I don't know if I have called the right place, but do you have...?" There must be one gateway to the organization and at the gateway must reside *one*, user-friendly tool, or knowledgeable helpful staff equipped with the tool, capable of directing the client to the appropriate sources. Different systems might underlie the one tool as long as a common look and feel is maintained.

The gateway may be replicated at different physical sites, but, again, it must have the same look and feel at each. It may be electronic and fully automated or supported by advisory staff. With regard to a Web site, caution must be exercised with regard to channelling the entrepreneurial spirit and constraining the egos which have seen "home pages" blossom as the vanity press of the electronic media. Each such initiative should be questioned in terms of what it costs to build and maintain and how effectively it contributes to the client's search. We must avoid the pitfall of building stove-pipe solutions to stove-pipe problems.

On-line, real-time

In a very short period of time the Web has significantly raised our expectations in our quest for information. We are satisfied with nothing less than instant, electronic gratification. While the Web is perfectly positioned to assist the client in browsing meta information, the question arises as to how to deliver a real product or service when the client finds something he or she wants. Clients now are less satisfied with generic products as we have seen the evolution of niche markets in which clients demand custom output suited specifically to their needs.

Once a client has been directed to a source of interest, it is in the client's interest and the agency's interest to provide the client with the facility to down-load, on-line, in real-time that information or data sought. The client's interest is obvious but the agency's interest is served in not only happy clients but also in hard cost reduction. The greater the capacity for a client to browse, specify, code or download, the less resources consumed by the agency. The technology exists to allow clients to download from public use micro data files and be billed automatically. Only in the case of confidential master retrieval files (which must remain behind fire-walls and screened for residual disclosure) is there a need to distance the client from the data. But even then, there is no reason why the client cannot code the request from record layouts, submit the job and have the agency produce the output and do the necessary disclosure screening.

With regard to the client who does not have the skill or the time to download his or her own data and information the option should be provided for account executives, using the same tools, to custom-build outputs to meet the client's niche needs. As the meta information opens the data archives to the world it might also be expected that opportunities will develop for private sector consultants to undertake browsing, downloading and analysis on behalf of clients.

Addressing the disharmonies

It is unrealistic to think that all disharmonies can be eliminated between sources. Differences in methodology such as whether a question is asked on the doorstep, over the telephone or on a self-completed form may yield subtle differences in output. Nevertheless, most serious disharmonies can be eliminated with concerted effort. Meta information must also underlay any attempt at harmonization since it is only with a corporate inventory of data holdings and documentation in place that the disharmonies are fully revealed. The meta information can also become a model of best practices and even a template for the development of standardized documentation ranging from mnemonics used in record layouts to classification systems to definitions. The adoption of templates and standards also promises the potential of hard cost reduction as future sources are developed. There is, however, no avoidance of the discussion and negotiation that must take place between the source areas with a view to the development of those standards. And there must be a commitment to eliminate the disharmonies.

Increased thematic output

The integration of data in a thematic way will also be facilitated by the construction of meta information. In the past, analysts may not have known of many relevant sources which existed, but armed with appropriate meta information, search tools and retrieval systems there is no reason why all relevant data cannot be ported to the desk-top. It remains, however, for the analyst to understand the importance of integration. At least, aggregated or tabular output should be accompanied with pointers to other related sources. At best, analytical or descriptive output should incorporate *all* relevant data and information in the analysis or discussion. It must be realized that the release of anything less than our comprehensive knowledge of an issue or population is as potentially damaging to our clients as are undetected response or processing errors. It is indeed curious that the statistician who shows such a proclivity for footnotes on methodological issues should have been so silent with regard to other sources of information or data relevant to the client.

Corporate Initiative

The question remains whether the above-noted steps can be undertaken without corporate initiative. As long as the corporate culture is such that it rewards individual production rather than corporate production it is doubtful that change will happen. Unless the stove-pipe production areas perceive some advantage in improving whatever performance measures against which they are evaluated they are unlikely to take initiatives. Perhaps some will, creating a groundswell in which others must join or be left behind. Even so, is there not too much at

stake to leave such developments to random individual acts? Is there not the possibility of duplicated effort and wasted resources? Does the lack of a shared vision, strategic planning, direction and funding from the corporation send the signal that integration is not really a high and urgent priority?

Information technology today presents unique challenges and opportunities to statistical agencies but to seize them it will be necessary to place a high priority on integration. That suggests the establishment and funding of a centralized body within the organization charged with leading the above-noted activities.

4. Conclusion

The past

The organization of statistical information has been driven primarily by methodology rather than thematic content. The integration of data on the basis of issues, populations or geography, and attempts to convert those data to information, have been hindered by the structure of the silos in which they have been collected and archived. There has not been a corporate, or for that matter, client view of the richness and comprehensiveness of the data holdings.

The future

In the statistician's ideal world there would probably be complete record linkage between all sources of data and, as a result, full integration and harmonization. Few, if any, agencies, however, operate in societies that would tolerate such a manipulation of private information. The challenge, and the opportunity, therefore, lies in moving to corporate rather than consortium data management. Meta information, harmonization and thematic integration are imperative if we are to progress in moving data to information. Agencies which fail to accept the challenge and the opportunity provided by information technology will be quickly perceived as unhelpful and irrelevant.

5. References

Bradley, B. (1994) Metadata Matters: Standardizing Metadata for Improved Management and Delivery in National Information Systems. Discussion Paper. Ottawa, Health Canada.

Hammer, M. and Champy, J. (1993). *Reengineering the*

Corporation. New York: Harper-Collins.

Nordbotten, S. (1993). Unpublished paper. The Statistical Meta Information System Workshop. Luxembourg: Eurostat.

Probst, S. (1995). Keynote Address, Data Warehouse Symposium. Ottawa: Tanning Technology Corporation.

Tapscott, D. and Caston, A. (1994). *Paradigm Shift: The New Promise of Information Technology.* New York: McGraw-Hill.

Vozel, K. (1993). Technical Evolution White Paper. Discussion Paper. New York: AT&T.