IHSN

INTERNATIONAL HOUSEHOLD SURVEY NETWORK

The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency

A Case Study of Statistics Canada

Gordon Priest

IHSN Working Paper No 004
June 2010

www.ihsn.org

The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency: A Case Study of Statistics Canada

Gordon Priest

June 2010

IHSN Working Paper No. 004

Abstract

A brief history of statistical activity in Canada. The statistical agency loses credibility. The agency undergoes a reformation but demands for data integration continue. A political miss-step and the agency defends its independence. New demands arise for harmonized data, but the enterprise is unregulated and isolated islands of production persist. A parliamentary debate leads to crisis but the agency responds. Shifts in methods of data dissemination. A new directorate is committed to integration, harmonization and development of metadata. Building the metadata system. The last stove-pipe falls and metadata come of age. A new policy on standards emerges. A new household survey strategy is developed. External influences lead to eventual success. Lessons learned. Challenges to agencies in developing countries.

About the Author

Gordon Priest was born and educated in Vancouver, Canada. After spending 10 years in the labour force he enrolled at Simon Fraser University at the age of 28. Graduating five years later with a Master of Arts degree, he joined the staff of Statistics Canada in Ottawa. His first assignment was writing the administrative report for the 1966 Census. Subsequent assignments related to both the planning and analysis of the housing, household, family and social content of the Canadian quinquennial censuses. Positions held included Acting Director of the Census Operations Division, Acting Director of the Demography Division, Acting Director-General of the Census and Demography Branch, as well as Director of the Social Statistics Division and the Director of Integration and Development of Social Statistics..

Over time, new duties in addition to the census work were added; these included the conduct of an experimental Family History Survey that subsequently morphed into the now regularly-conducted General Social Survey. During the same period his division launched the well-respected journal Canadian Social Trends, which remains one of the agency's flagship publications. The division also began to produce thematic reports of integrated data on a wide range of target groups, including the aged, children, young people, women and lone-parent families. Customised analyses and reports were also produced for other government departments.

Having become a fierce advocate for data integration and harmonization, his last assignment was Director of Integration and Harmonization for Social Statistics. The mandate included both building a metadata base of social statistics within Statistics Canada and working with agencies such as the Conference of European Statisticians and Eurostat in promoting international standards.

Long a supporter of volunteer work, Gordon has served on the Board of Directors of the Crohn's and Colitis Foundation of Canada, the board of directors of a housing consortium and as a mentor at an immigrant settlement centre. In retirement, Gordon, his wife Barbara (and their Maine Coon cat Cassie), parents and grandparents to two daughters and three adult grandchildren, patiently await the arrival of great grandchildren!

Acknowledgments

The document was developed for the International Household Survey Network (IHSN) with financial support from the World Bank Development Grant Facility, Grant No 4001009-06, administered by the PARIS21 Secretariat.

It was prepared by Gordon E. Priest with contributions, in the form of input or comments and suggestions, from Ernie Boyko (consultant, IHSN), Olivier Dupriez (World Bank), François Fonteneau (OECD-PARIS21). Rosemary Bender and other staff members from Statistics Canada commented on the paper and provided many helpful comments. The findings and views expressed in the paper may however not be considered as representing those of Statistics Canada. John Wright edited the document, and typesetting was by Rhommell Rico.

This paper (or a revised copy of it) is available on the web-site of the International Household Survey Network at www.ihsn.org.

Citation

Priest, G., 2010. "The Struggle for Integration and Harmonization of Social Statistics in a Statistical Agency: A Case Study of Statistics Canada", International Household Survey Network, IHSN Working Paper No 004, June 2010.

The findings, interpretation and views expressed in this paper are those of the author and do not necessarily represent those of the International Household Survey Network member agencies or secretariat.

Table of Contents

Introduction

Canada is a large and diverse country with a small but equally diverse population, most of whom live within 100 miles of the southern border. When Europeans first settled permanently in what is now Canada, the already resident aboriginal people spoke over 50 different languages grouped into about 12 major linguistic families. Although Canada has both English and French as official languages, over 20% of the population reports neither as its mother tongue. Nearly every language known on earth is reported by at least some Canadians. Canadians also have diverse ethnic and cultural backgrounds. The country has always had a relatively large immigrant population, contributing further to its diversity. Canada also has a very large and diverse geography: from the frozen north to the temperate south; from soaring mountains to plains and lowlands; and from the Pacific to the Arctic to the Atlantic oceans. Neither its geography nor its population is easy to enumerate.

Nevertheless, this diverse country has been well served by statistical services that date back over 400 years. Statistical data have served demographic, social and economic planning throughout both French and English colonial periods – from the early confederation period, when there were only four provinces, to modern times, with 10 provinces as well as three territories. The national statistical agency, Statistics Canada – through the quality and objectivity of its work and statistical output, and its contribution to international work, such as the Conference of European Statisticians and Eurostat – has gained a world-wide reputation for excellence. Indeed, for a number of years, the *The Economist* magazine rated Statistics Canada the best statistical agency in the world.

Notwithstanding such laurels – as we shall see in the following account of the struggle for integration and harmonization of social statistics – as the agency began to expand its statistics' programs many decades ago, there were many significant inefficiencies, internal competition rather than collaboration, under-utilized data and misleading outputs due to production-segmented data releases that did not provide a comprehensive view of a situation. Though Statistics Canada has made great strides to address these issues, ensuring integration and harmonization of social statistics is, in fact, an ongoing challenge with an ever-evolving society, new technology and the development of new data sources. National statistical agencies that fail to learn the lessons herein and address matters of integration and harmonization are destined to repeat the same mistakes.

Flaws in Statistical Series Revealed

It was late-1979 and heavy dark clouds scudded across the sky. Rain lashed the windows of a conference room of the national statistical agency. Inside, the three directors sat silently, each deep in thought concerning the meeting about to take place. A clerk arrived with a coffee trolley and the rattling cups further jangled the nerves of the seated executives.

At precisely 2:00 p.m. there was a knock on the door and a senior member of the Chief Statistician's staff entered, followed by a well-dressed woman and a TV cameraman. Introductions were made as the cameraman went about setting up his equipment. The woman was a well-known and well-respected journalist for one of the national networks. Of late, she had been highly critical of the agency's media relations and the validity of some of its statistical concepts and methodology.

This was the environment in which the three directors found themselves on the day in question. Their assignment was to answer the journalist's demand for an explanation of why the agency had recently released three different counts of the number of families in the country. There had been no time for a dress rehearsal and under the glare of the TV lights the directors were clearly uncomfortable. Each tried to explain the methodological reasons for the differences and, unfortunately, each tried to protect his own domain. Each suggested his or her series was the most accurate, the most reliable or the timeliest. Initially, it seemed the journalist had her story: it appeared the three 'knaves' were going to bring down a house of cards. They used terms like *sampling*, *administrative records*, *derivation algorithms*, *imputation* and *random rounding* to defend their figures. The journalist became agitated, sensing she was being buried under a landslide of statistical jargon. She said she had heard nothing she could clearly relate to her audience that would help them understand the different numbers.

At this point, one of the directors rose and approached the blackboard, picked up a piece of chalk and drew a circle. Inside the circle he drew a dot just inside the two o'clock position. He asked the journalist what she thought it represented. She said it looked like a bubble or maybe a balloon. He drew another circle, put a dot in the centre and drew lines radiating to the circumference. Again he asked what it represented. She replied, "A wagon wheel". He drew a third circle, this time with lines from the top and bottom poles –

not unlike lines of longitude on a globe. The journalist said that circle looked like a globe or maybe a ball. The statistician said her answers were imaginative but what he intended to represent were three oranges. The first was unpeeled and the dot to the upper right was the navel. The second was an orange that, with the navel at the top, had been cut at the equator, revealing two halves and all the segments – in a manner a grapefruit is usually served. The third drawing was also an orange – intact but peeled to reveal the segments in a longitudinal view.

The statistician said the three oranges were not unlike the three series of data on the family. The three drawings all represented an orange but from a different perspective. He said the three data series all represented the family but from a different perspective. They all showed different characteristics of the family and each series had evolved to satisfy the needs of different clients. He suggested each series revealed important information about the family and each series as it stood was legitimate. He added it was unfortunate that the totals did not agree but perhaps that was less important than the understanding of family dynamics.

The journalist told her colleague to turn off his camera. She sighed and then said, “Well, I don’t have the story I thought I had. I now understand the rationale for the three series”. The directors relaxed. “But”, she added, “I have another story and I shall certainly pursue it. For a group of people who have the responsibility for taking the pulse of the nation you are certainly out of touch with it. You are so focused on your own little section of the shop floor and on your own narrow client groups that you have no idea there are other client groups out there who need integrated and harmonized data; and if you can’t manage integration then you need to clearly explain the differences. What we have here is a sorry case of failure to communicate with the Canadian people and I don’t know whether it is because you are introverted or arrogant.”

A Brief History of Statistics Canada

Canada has had a strong statistical program going back to the French regime when missionaries in the colony first made a population count in 1605 and began to compile records of births, deaths and marriages. The first census was conducted in 1655 to support demographic, social and economic planning in the struggling colony. When the British took over the French territory in 1763 the censuses and statistical programs continued intermittently until 1851 when regular decennial censuses were initiated. The practice was continued after the confederation of Canada in 1867.

Rebellions in the western regions in the late 1800s precipitated a demand for supplemental regional censuses. A profound growth in the west led to the formation of new provinces and the demand for regional censuses on a five-year cycle rather than every 10 years: a practice extended to all the country in 1956. The depression of the 1930s and urban civil unrest led to the initiation of neighbourhood (initially called social areas and subsequently census tracts) data in 1941. World War II and the post-war baby boom and urbanization led to further demands to expand data collection activities.

Although the *Census and Statistics Act*, passed in 1905, saw the centralization of statistical activities in a permanent Census and Statistics Office, other federal government departments and some provincial governments continued with their own statistical programs. By 1912, it was recognized that, overall, the statistical system was in disarray. An investigation was launched (the Foster Commission) and it found that official statistics were “fragmented and poorly coordinated”.

The Commission found in part:

> *Though many of the statistical reports issued by various departments and branches are of undoubted excellence…there is…a lack of coherence and common purpose. This is traceable to imperfect appreciation… that the statistics of a country…should constitute a single harmonious system…* [25]

In 1918, the Census and Statistics Office, reflecting its broadening mandate, was renamed the Dominion Bureau of Statistics, which it remained until 1971 when a new Statistics Act changed the name to Statistics Canada. While most statistical branches of federal government departments were transferred to the central statistical agency in 1918, divisions were established internally that reflected in part methodology and in part subject-matter interests. For example, based on different methodology, vital statistics, utilizing administrative records, were collected in a different unit from the census. On the other hand, subject-matter divisions saw different units responsible for immigration and emigration data, education data, income and wealth data and criminal data. By the mid-1940s the national agency had embraced sampling methodology and introduced a labour force survey, yet another source of statistics. As methodological sources of data increased so did the production of data on particular subjects or populations: demography, housing, households and families, health, education, labour, income, justice and crime, and social attributes such as language use, ethnicity, immigration status and citizenship. Population estimates and projections, integrating administrative and census data, were introduced as both new methodology and another new source.

By 1944 it was evident that both methodological and subject-matter units within the Bureau had become independent producers, negating many of the gains made in centralizing the statistical program. An officer wrote to the Dominion Statistician of the day stating that, "There appears to be an immediate need for a study of...classifications...even in the Bureau classifications are not uniform...each Chief being a law unto himself." [25]

The 1970s and 1980s saw a further extension of survey activity with new surveys being introduced: for example the Household Facilities and Equipment Survey, the Canada Health Survey and the General Social Survey. A Special Surveys Division was formed to conduct custom surveys for specialized clients across a broad spectrum of interests and populations and subjects including aboriginal peoples, health and activities' limitations, graduates, child-care, work injuries, alcohol and drug use, smoking, volunteer activity, work history, violence against women and ageing and independence. The good news was that the range of data on issues on and relevance to a wide range of clients increased dramatically. The bad news was that this rapid expansion of activities led to the potential for further disharmony to creep into the statistical series.

The Agency Under Siege

At this point it is necessary to retrace our steps somewhat to events that led to changes in the corporate culture. The 1971 Census, using new processing technology, was plagued by both conceptual and technological problems. Products were very late being released and some statistical series were cancelled due to data errors. Confidence in the agency was eroded. Furthermore, in 1978 and 1979, difficult economic times led the government to pressure the agency into making significant cuts to both the agency's products and programs. In 1978 alone, 96 publications were eliminated and a further 28 reduced in content**.** [25] Staff were downsized and those who remained largely lost their direction. Public trust in the agency was clearly eroded as both clients and media became frustrated with the agency, and relationships deteriorated steadily. At one point, staff were forbidden to speak to the media on any issue and all questions were forwarded to the office of the Chief Statistician. That office alone could not handle the volume of inquiries and thus communication essentially broke down. A siege mentality evolved and the media circled constantly looking for another example of a government department in chaos.

It was in this environment that the journalist cited in the opening paragraphs was granted access to the three executives. The views she expressed at that time were largely reflected in an independent inquiry into the agency by Sir Claus Moser (the Moser Report), former Director of the United Kingdom Central Statistical Office, and in a study of management practices by Price Waterhouse Associates. The investigations showed that public perception of Statistics Canada was not flattering. The public viewed the agency as a rather cold and aloof organization housed in a remote concrete bunker, staffed by cadres of dull people wearing bow ties and thick glasses. They saw these automatons hunkered down over their desks, tied to computers, or maybe even quill pens, cranking out endless tomes of numbers – numbers of interest only to government bureaucrats, academics or multinational corporations. Clearly, things had to change.

The Reformation

In the past, decisions about information to be collected were made primarily within relatively narrow parameters dictated by the needs of federal government departments and provincial governments. Following the inquiries the process of consultation was broadened significantly. Regular (sometimes as much as once a month) bilateral consultations on data needs with senior officials of federal government departments were instituted. Although there had been long-standing regular meetings, under the auspices of a federal-provincial agreement, they were revitalized to ensure the data needs of provincial and municipal governments were addressed. Periodic consultations were initiated with non-governmental organizations, such as labour unions, professional societies and advocacy groups. Employees of Statistics Canada were encouraged to participate in the meetings and activities of learned societies such as the Canadian Population

Society, the Canadian Economics Association, the Canadian Association on Gerontology and the Canadian Sociology and Anthropology Association.

A further element in the consultation process was the formation of a National Statistics Council that shared… the common purpose of providing high level policy guidance to the statistical system and of serving as yet another protection against politicization. Its members include eminent people from business, universities, research institutions, provincial government, labour unions*, the media…but not the federal government. While their explicit mandate is to advise the Chief Statistician about broad policies and priorities, members of the Council are appointed by the Minister responsible for Statistics Canada and have therefore access to him should they think that the agency is threatened…either because of political intervention or lack of adequate funding. The Council's very significant influence and standing derives from the eminence of its members.* [9]

Professional advisory committees were also founded, operating *in such diverse fields as demography, social conditions, health, agriculture, service industries, price measurements, science and technology statistics. Their membership is selected on the basis of individual expertise without regard to issues of 'representation'. Their role is to challenge the status quo in terms of both content and broad methodology. Typically they meet twice a year for two days, and members serve without remuneration. The head of the substantive program most directly concerned serves as the secretary, but meetings are widely attended by staff. Most committees' contribution is channelled through the informal interactions provided by this arrangement, but*…written recommendations to the Chief Statistician are encouraged. [9]

The renewed consultations with federal departments, the National Statistics Council and the professional advisory committees all proved useful in terms of addressing difficulties in mining data from administrative records. While, if given the will and given the tools, the agency could begin to address problems of metadata and lack of harmonization within its own collection vehicles, administrative sources from jurisdictions over which the agency had no control were intimidating. In fact, at the confederation of Canada responsibility for health, education and vital statistics was delegated to the provincial governments. Attempts to integrate data from so many different agencies that often had different mandates and different systems of record-keeping and used different technology had always been a formidable challenge. The consultations with supplying jurisdictions needed to move forward and this work was greatly aided by both the NSC and the professional committees, as they all comprised capable and well-respected members of the community with considerable influence.

Not to be overlooked was the unprecedented agreement between the government and the agency that senior staff, on a highly-confidential basis and sworn under the Official Secrets Act, would be given access to secret cabinet documents setting out government intentions. This served the dual purpose of ensuring that any data used in the documents were the most appropriate and correctly interpreted and allowing the agency better to anticipate data or information needs that might be precipitated by new policies or programs.

Responses to Demands for Data Integration

Through the initiation of all these consultations, a message that came through loud and clear was that clients needed not just tabular data outputs but access to actual data such as public-use microdata files, as well as analyzed information based on integrated and harmonized inputs. This required another significant cultural shift for the agency and its staff, who were sent on media training and taught how to produce interesting but factual reports aimed not at government bureaucrats or academic journals but at ordinary people who read morning newspapers and listened to the evening news. The major change in policy from releasing simple data to releasing analytical reports that reflected trends and relationships added significant value to the outputs. The integration and analysis of data on specific issues, specific populations and specific geographical regions immediately led to a breakthrough in getting accurate information into the hands of all Canadians. In 1993 the agency began to track the degree to which its outputs were used in the media. In one day alone the print media featured agency releases in reports on workplace injuries, bilingualism, smuggling, shopping patterns, violence against women, gender differences in educational attainment and job promotion, and school enrolment of teenage mothers. [15] In the same year the analyzed results of a special survey on violence against women were released. A sensitive topic that had previously been discussed primarily behind closed doors and shuttered windows became, within two weeks of release, a major media story – and Canadians became not only informed but began to participate in public discussion. [15] The public debate led to political debate, which in turn led to the development of new programs that better protected women.

The prime vehicle for new data releases from the agency had long been a publication entitled *The Daily*, which was the official release mechanism of the agency. It was released at exactly the same time every day – a time that best accommodated people living in the country's six time-zones. A change in policy dictated that whenever a major release was due, such as household and family data from the census, two lock-ups were held two hours before the official release. One was for government officials whose mandate covered the subject of the release. The other was for the nation's media. In addition to copies of *The Daily* and its new analytical and informative content, briefing notes and press releases were provided detailing the more significant findings. No one was allowed to leave the lock-ups until the time of official release, at which point government ministers and staff scurried back to their offices to prepare to answer questions in Parliament or prepare statements for their own constituents. The urgency of their work depended very much on whether the released information reflected positively or negatively on their departmental policies and programs!

The second lock-up was for members of the print, radio and television media. Media training was provided for agency staff. They were taught how to produce objective, analyzed outputs that could be quoted directly in the media. Journalists, generally not trained in statistical analysis, no longer had to attempt to integrate and analyze data. Often the media now used the releases from *The Daily* and other sources verbatim in their news reports and sent journalists and camera crews to interview experts on the subject or ordinary people to put a 'human face' on the newly-released information. Furthermore, errors of statistical interpretation in media stories reduced dramatically. This was a 'win-win' situation: the journalists' workload was reduced and the agency had more confidence that their outputs were being used in an unbiased, correct and objective manner.

Another initiative was the launching of a quarterly journal entitled *Canadian Social Trends* – a landmark publication in integrating data from diverse sources on topical issues such as women in male-dominated occupations, the decline of unpaid family work, wife abuse, employment of disabled persons, seniors, immigrants and household-shelter costs. The publication quickly became a 'best-seller' among the agency's publications. It was displayed prominently in public libraries; became required reading in schools and universities and a reference document for both government departments and non-governmental agencies; and was often quoted in the media.

At the same time, a new project was launched to undertake both custom research for other government departments and produce special publications on target groups. Integrated data from diverse sources were analyzed to produce profiles of special populations such as lone-parent families, immigrants, aboriginal people, children and seniors.

Other publications introduced or revitalized during this period were *The Canadian Economic Observer*, *Perspectives on Labour and Income*, *Canada: a Portrait, The Canada Yearbook* and the census-based series *Focus on Canada*.

A further innovation was the establishment of subject-matter committees with a view to reducing disharmony across databases and seeking opportunities to integrate data from the various sources. One example was the committee on families. The agency produced data on two different definitions of the family. One was called the Census Family, which was essentially the nuclear family. The other was called the Economic Family, which was the extended family. The committee was struck with a view to developing a standard based on one concept or the other. Failing agreement, then steps were to be taken to ensure that clients clearly understood the difference between the two concepts. Representatives from each of the divisions sat on the committee, which initially met once a month and reported periodically to a director-general. Some representatives took the matter seriously and worked hard at drafting proposals. Some representatives, possibly directed by their managers, simply blocked any progress. Reasons given included the need to maintain historical comparability, stating their clients would not accept change, or they did not have room in their publications to include footnotes explaining differences between the concepts. It was clear early on that in some quarters there was no interest in compromise, neither was there any real pressure from management to resolve the issues. After a couple of years the committee just stopped meeting.

In the same way, most subject-matter committees had only marginal success. The most probable reason for failure was there was no firm message from the centre that the agency must function as a single enterprise, and that it must serve markets on a corporate rather than a fragmented basis. Generally, the committees were given no resources to execute programs of work. In fact, following a government directive to put the marketing of products on a cost-recovery basis, the old practice of independent 'islands' was even reinforced by making areas of production enter into division-based marketing contracts. Thus there was no real incentive to commit any resources to either integration or harmonization, as competition was increased and cooperation fell by the wayside.

Fallout from Political Interference

In 1984 a government that had been in opposition for many years was elected to Parliament with an overwhelming majority. This new government made no attempt to hide the belief that the bureaucracy had served the old government so long they were biased and could not be trusted. Even though Canada had a long-standing professional public service with appointments and promotions made on the basis of merit rather than political patronage, 'shadow' deputy ministers were politically appointed to major federal departments, including Statistics Canada. Work was hindered and programs delayed as these deputy ministers and their aides spied, pried and tried to uncover allegiances to the old government, which was now in opposition. Certainly, they were generally ignorant of statistical methods and even of the importance and use of statistics. Their questioning delayed work and their meddling introduced risk. Eventually, finding no political bias in the public servants, the appointees were called off, but not before leaving staff shaken and nervous.

However, the government had other axes to grind. It believed that services provided by government should have a market value. It sold many of the long-standing Crown corporations to private-sector enterprises. It introduced user fees and cost recovery for government services. Not only did the public have to pay for services and products but so did government departments for services from other government departments. Even internally, in Statistics Canada for example, divisions had to pay each other for data or other services provided. Divisions that were producers of data did stand to profit but divisions that were integrating data from other divisions were severely disadvantaged. The only legitimate way they could find the funds to pay for the data was to take it from budgets intended to purchase supplies, material or equipment. Personnel in the integrating divisions also became stressed and overworked when their roles changed from writers and editors to analysts as well. It was a field day for Financial Operations as their staff grew in order to manage the new financial systems needed to track the burgeoning transfers.

It was nearly disastrous for new programs of integration like *Canadian Social Trends* because the program (even though it did not have the budget to do so) had to purchase data from other divisions and had to pay for the time of analysts from other divisions who did research or wrote for the publication. Collapse was prevented only through the development of an underground black market in which analysts and data were bartered under the table. Some analysts also did work in their own time for the prestige of getting their work published in the journal. Another threat to the publication came from a dictate that publications had to be fully cost-recoverable. Due to market size, the English edition had no problem covering its costs, but the French edition lost money. An order was made to cease publication of the French edition. However, under the Constitution, the agency was obliged to produce the journal in both official languages. The order was then made to cease publication altogether. Only loud public protest from its readers saved the journal, which continues to be published today every six weeks.

Another decision of the government of the day had far-reaching implications for the agency and its major clients. The Prime Minister and a couple of cabinet colleagues, in reviewing budgets one evening, noted a large-ticket item. It was the 1986 Census. Without consulting other cabinet colleagues, deputy ministers, provincial governments or even the Chief Statistician, they announced the next day that they had cancelled the 1986 Census. In fact there were a number of pieces of legislation that required the Census to be taken. In order to follow through their decision they would have had to change all that legislation – an impossible task under the circumstances. They never did admit they had made a constitutional mistake but subsequently announced they had found ways of reducing the costs and were therefore able to allow the Census to proceed. The smoke-and-mirrors process by which the money was found was to charge federal government departments a share of Census costs. Other clients faced significantly-increased user fees and product costs. As we shall see subsequently, these kinds of government measures continued to dog attempts to improve integration and harmonization.

Need to Maintain Political Independence

Statistical programs have often been defended on the basis of three needs: the need for informed government and corporate decision making; the need for responsible fiscal sharing of resources; and the need for representative government. Certainly, the release of information rather than raw data supports the first two but it also supports the third in ensuring that the electorate is well-informed on issues on which governments is, or should be, developing policies and programs. It also provides a report card on the effectiveness of those same policies and programs. An independent, credible, scientific and objective statistical agency that consults with a broad cross-section of its clientele and produces easily-understood information is the most effective method of ensuring a well-informed electorate that can hold its government and its officers accountable. It is ironical that about the time that the agency was dealing with the threatened cancellation of the 1986 Census they were also providing assistance to another country in analyzing data from its most recent census. The minister responsible for statistical programs in that country told the Statistics Canada delegates that it was absolutely essential that statistical activity was honest, unbiased and objective, and that outputs must be easily understood, not only by the country's leaders but its populace as well. She said that was something very precious that must be carefully nurtured and guarded. She added that in her country, census results had been corrupted by political leaders for political ends, with the result that economic, social and demographic programs had failed because the planners had not known they were working with corrupted data.

New Demands for Harmonized Data

The move from production of independent data-series to thematic information saw the agency becoming a user of its own data. In this role of integrating data from different sources it was quickly realized just how disharmonized the data from different sources were. This problem was not unique to Statistics Canada. During this period, the UK Central Statistical Office was also producing a publication called *Social Trends* in which data and information from various sources were presented. The editor at the time, in discussion with Statistics Canada staff, noted that he had the same problems with independent producers.

As in the UK, officials of Statistics Canada responsible for these integrated data programs began to lobby loudly for the agency to clean up its act. For example, a request had been made for the agency to produce a comprehensive report on seniors: showing the type of household they were in; whether they were living alone; whether they required support or were fully independent; whether they were home-owners, tenants or in an institution; their income by source; expenditure on shelter; health status; and their social activities. It called on sources as diverse as administrative records, the Census and a significant number of regular or one-off household surveys. Frequently, the count of the population in question did not agree from source to source, neither did definitions nor classifications of variables. It was extremely difficult to prepare a comprehensive profile of the population in question without compromising the clarity of the report with copious caveats and footnotes – more statistical jargon guaranteed to turn off not only the Canadian public but also decision-makers.

Even more difficult, however, was the task of persuading any of the independent producers of the data series to share their data in this corporate undertaking. Many had a proprietary view of 'their' data, stating that it should only be released in 'their' publications. Others, seeing the success of the new flagship publications, sought resources to produce their own competing journals. Further, and probably without exception, they argued that their first duty was to their own long-standing clients and that maintaining historical comparability precluded integration and harmonization. The counter argument was that their product would be significantly more useful if it could be integrated with other sources, but to reach that goal compromises would be needed to develop standardized and harmonized concepts, definitions and classifications. It was argued that all clients would be better served in the longer term by integrated and harmonized outputs. The historical comparability argument is always more difficult to counter, especially for those whose mandate is to study long-term trends or make future projections. That is a legitimate concern and a tough nut to crack. Nevertheless, many statistical series have had to be interrupted due to various external forces. For example, in the 1980s the agency had to abandon its concept of head of household when public and subsequently political pressure demanded that it was no longer a relevant concept and the characteristics of such a person should no longer be produced. Unfortunately, the concept, in conjunction with other characteristics such as age, gender and marital status, was used to derive family statistics and a new concept of Person One had to be introduced solely for the process of the derivation of families.

In 1982 the passage of the *Constitution Act* and its reference to aboriginal peoples as Indian, Metis and Inuit, and the subsequent Royal Commission on Aboriginal Peoples, meant the agency had to develop new concepts and classifications that largely fractured any past historical comparability. More recently, both public pressure and some legislated changes on the issue of same-sex unions or marriages, have forced some dislocation of the concept of family. Therefore, much as we wish to maintain historical comparability, change is inevitable and the case for refusing standardization and harmonization is weakened.

An Unregulated Enterprise All the Same?

In spite of arguments like those above that historical comparability might well be compromised anyway, directors of the various independent data sources still resisted change. The agency was in many respects an 'unregulated enterprise', as described by Tapscott and Caston in their book *Paradigm Shift: the New Promise of Information Technology*. [27] They describe islands of technology or expertise that meet specific needs but result in 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, they become balkanized with formidable physical and organizational barriers, redundancies and inefficiencies. They state: *"Lack of integration and gaps between systems islands also caused miscommunications and lost opportunities to achieve business value…Operations and customer service were restricted."* With regard to customer service, they used the example of the banking industry where customers were shuffled between the savings department, the mortgage department, the loans department or the credit card division. In effect, the relationship between the organization and the client did not focus on the client.

This level of client service was no longer acceptable. Organizations of the 21st century need to function as a single enterprise rather than a collection of business units. The new enterprise, suggest Tapscott and Caston, must be integrated with an overall strategy and architecture for the business, work organization, information and technology. Furthermore, the isolated technological applications of an earlier time are no longer adequate. "*Companies are discovering that they have to establish enterprise capabilities that will create new opportunities for sharing and reusing information and information technology…more and more organizations are becoming aware that the technical and structural barriers that have previously prevented or hindered internal communication and the sharing of resources must be dismantled. There is a growing need for direct links between sources of information and the people who use it…*" [27]

In the course of human history there have been discoveries or inventions that have irrevocably changed our collective development: the control of fire, the development of agriculture, the development of trade, the development of the wheel, the industrial revolution and now the information revolution. Tapscott and Caston argue that we are already entering the second era of the information age.

The first era really began in the 1950s with the introduction of mainframe computers to the management of organizations. Early applications were in the management and control of physical assets and facilities, financial management and control systems, and the management and support of human resources. In the case of statistical agencies, there was also the application to the capture, processing, storage and retrieval of data as a product. The result was the development of islands of technology that were rigidly and centrally controlled and which served a relatively small number of technocrats or bureaucrats. Most users or potential users were marginalized.

As early as the 1970s, some federal government departments complained that it was difficult to deal with the agency because they had to deal with a multitude of players, and comparability was often lacking between their data products. Subsequently, in the early 1980s, an attempt was made at developing a harmonized set of housing, household and family definitions. [24] Despite reporting to the Conference of European Statisticians on attempts being made within the agency, the efforts towards harmonization largely failed due to a lack of cooperation. Similarly, an attempt to cross-reference the availability of related sources of data within the agency also failed as the producers saw themselves as competitors. An early attempt to build a metadata base on housing, household and family data sources also failed, in part because of the lack of cooperation but also because the agency saw no future in the initiative and failed to grant funds for its development.

Meanwhile, the initiatives of the '70s and '80s related to the development of social indicators blossomed and waned, not only within Statistics Canada but also internationally. Certainly, part of the problem was centred on the inability of statistical agencies to develop any collaboration between the islands that would feed a system of social accounts. But there was also the overwhelming conceptual problem of how diverse data based on diverse universes and diverse sources could be integrated and weighted in some way to develop either a single aggregated indicator or a series of complementary indicators. Furthermore, the difficulties were compounded by the lack of a framework for social reporting and a lack of consensus with regard to the meaning of social indicators. What did follow, however, while perhaps not the social indicators that had been envisioned, was the development of social reports comprising descriptive statistics that some have argued have little ability to explain causal relationships or offer any predictive power.

In fact, Carley argued that, "There is little evidence that social reports are used to any great extent by decision-makers, except perhaps as general background data." [2] That view was confirmed to some degree by discussions between Statistics Canada staff and policy analysts associated with federal government departments. What must not be overlooked, however, is the degree to which social reports, when widely reported in the media, form the basis for public discussion. That very public discussion, if elevated to a sufficient level, may precipitate parliamentary debate, which in turn may lead to the call for policy formulation or policy review at the least and program initiation or review at the most. The real value in social reporting, as it has evolved, has been the degree to which it has supported informed public discussion. As such, it has been an agent of social change in itself. [15]

Isolated Islands of Production

In Statistics Canada, which had become both a user and a producer of data, the new technology noted by Tapscott and Caston was relatively quickly embraced and used in the production of information for external clients. Unfortunately within the agency there were already the well-established islands of expertise. These islands were centred on both subject matter divisions and methodology. That is, on the census activity, household survey activities and administrative record activities. Competition rather than cooperation was more often the practice, and when the new technology became available different tools and systems were developed rather than integrated ones.

Goals were specific to each of the independent statistical divisions and each sought to develop its own supporters and clientele. Overlapped lines of inquiry began to appear across the various collection vehicles but little attention was paid to the development of standardized concepts, definitions and classification systems. There was little communication between staff working on different vehicles and often staff on one vehicle did not even know about similar enquiries being made by others. In effect, the islands of technology and expertise had become balkanized. Competition rather than cooperation was the method of operation.

Not only was the full potential of outputs and products not realized, but staff were used inefficiently. Clerical, technical and professional personnel were all classified into highly-specific jobs related to particular technical skills or subject-matter expertise. Because of the cyclical nature of surveys and censuses there were times when some individuals had too much work and at other times were not fully utilized. In fact, Canada has a legislated requirement for a five-year census cycle (which very well serves small geographical areas and small populations). There are times in the cycle when some individuals might be working on different phases of three censuses at the same time. The census used a project-management system and each census cycle had a different manager. Generally, they refused to coordinate their schedules and their demands on key staff with the result that some people suffered at times from impossible workloads, which in turn led to significant stress and burn-out. There were periods when a shop suffering from excessive demand would ask to borrow staff from another area. The other managers, not wanting to admit their people were underutilized, would invent projects rather than loan staff to other areas. The system was competitive, highly inefficient and demoralizing, both for many staff and for managers.

Eventually, an initiative was taken by some managers to develop generic job descriptions that comprised a broader range of tasks, skills or professional knowledge. Initially, there was considerable resistance from some managers who could not see the advantages. There was also resistance from some staff, their unions and even from the Public Service Commission that had overall responsibility for staffing in the public service. It was a long and difficult battle to change the corporate culture but slowly progress was made. Many clerical and technical personnel began to work out of 'pools' from which they could be assigned to areas of need. They began to realize that there were positive benefits in that they learned new skills or expertise, becoming more knowledgeable and valuable employees. It reached the point where employees began to seek such assignments in order to advance their careers. A program was developed to assist personnel in finding temporary assignments and it was so successful that professional employees joined in and, as success led to success, the program was even extended to other departments. It was a 'win-win' situation with unforeseen corporate benefits – it was also an important factor in helping to break down barriers between the isolated islands of production.

Keith Vozel of AT&T, in his *Technical Evolution White Paper*, [29] described such 'island' organizations as vertical or stove-pipe, the parts of which tended to address a single issue or client without regard 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. They represent untapped potential and lost opportunities.

Many of the agency's clients expressed much the same views. Officers of Health Canada, who had done much work in the field of metadata development, described Statistics Canada as an organization of 'autonomous data programs'. In the early 1970s, at a meeting of the Conference of European Statisticians, called to begin discussions for the 1980 round of censuses, the delegate from Statistics Canada had dinner with one from another Canadian federal agency. The latter noted that it was strange that Statistics Canada should be supporting international integration and harmonization of data when its own data holdings were in such disarray. He noted that he had to go to as many as eight divisions in the agency to obtain the data he needed and then often found there was little comparability between the sources. [17] Here again was evidence that while many clients were well-served by individual producers, those who needed data from multiple sources were not.

By the 1990s, new technology in personal computing had precipitated a paradigm shift. A new generation of clients, emboldened by the power of the Internet, developed new expectations, particularly with respect to the search for information. These clients, all with their unique and particular needs, expected to be able to browse meta information thematically, determine sources, make selections and even download: on-line, real-time seamlessly, and at little or no cost. The soft underbelly of statistical agencies was revealed as they were in no position to respond.

Nordbotten, addressing a Eurostat workshop in 1993, noted: [13] "*Users have little knowledge about the content of statistical data archives, how to combine statistics from different sources and how they could benefit from the large sources of potential information hidden in the data archives. To the extent that producers themselves know the content of their own statistical data sources, the necessary keys to open up the treasures properly for the users are not implemented. These keys are the statistical meta-data systems.*"

Parliamentary Debate Leads to Crisis

Nordbotten could have been addressing his remarks directly to Statistics Canada as it was the very next year that there was public discussion and parliamentary debate about immigration issues in the country. The federal department then responsible for policy and programs was Citizenship and Immigration Canada (CIC). They were accused of not knowing the impact of immigration on the nation or knowing how well immigrants fared after arriving in the country. In their defence they argued that there were simply insufficient data available for the necessary research and the spotlight fell on Statistics Canada. A call went out to the 'island' producers within the agency to identify data series that would be useful to the immigration researchers. It was immediately clear that the officers in the agency did not immediately know the extent of their holdings of potentially-useful data, for there was no systematic inventory of them.. After an essentially manual search, which took some weeks to complete, it was revealed that the agency had a significant inventory of data related to immigration issues. While the CIC officers were surprised by the amount of data that were revealed, they were no more surprised than Statistics Canada's own officers.

Within months of the CIC matter, senior officers of another department, Indian and Northern Affairs Canada, were required to prepare a briefing note for their minister on the social condition of Aboriginal Peoples. They again indicated they were unable to prepare a comprehensive report because of the lack of relevant data. They also expressed frustration that there was no one source to whom they could go to help them open the right doors in the agency. Again the spotlight focused on the agency and again a manual search of the archives indicated a considerable array of data that could be useful to the department.

At the same time, the National Council of Welfare, a non-governmental organization, wrote to the Assistant Chief Statistician responsible for social statistics to express. concern about the lack of harmonization in the classification of lone parent families. "*We urge you and your colleagues",* they wrote*,* "*to adopt standard classifications as soon as possible, and to use them in the 1996 Census and all your other publications*."[1] Here was clear evidence that the subject-matter committees, established more than a decade earlier, had had little or no success.

Again, according to Tapscott and Caston,[27] there is no place in for organizations that do not recognize the empowerment of their clientele. They say that the new era and new enterprise will be open and networked. "*It is modular and dynamic – based on interchangeable parts. It technologically empowers, distributing intelligence and decision-making to users. Yet, through standards, it is integrated, moving enterprises beyond the system islands (and their organizational equivalents)…It works like people do, integrating data, text, voice and image information in various formats…*"

The Agency Responds

With major clients in full cry, and with increasing demands from data users within the agency, changes had to be made. While the agency had long had a Standards Division, it was primarily occupied with economic statistics such as the Standard Occupational Classification and the Standard Industrial Classification, as well as with standard geographical classifications. Little attention had been paid to social statistics. The first step in the social statistics' field came when the director responsible for producing *Canadian Social Trends***,** who was one of the harshest internal critics of the agency's lack of progress on integration and harmonization, was assigned to the task of building a metadata base on sources of data related to immigration. The work involved a painstaking search of all data sources that might contain immigration-related data. Definitions and classifications and questionnaires for collection vehicles or derivation algorithms for administrative records for each source were identified and documented. The depth of the disharmony in the outputs that was revealed was profound. Even in rare cases where any two sources might have had the same definition and classification of immigrants, variables used in cross-classification were frequently different. For example, different age groupings might be used or different levels of educational attainment or different measures of labour-force activity. Therefore, attempts to integrate or compare the data from the different sources were severely limited and the potential value of the data was severely compromised. In agricultural terms it was like building one tractor to pull a plough, another to pull the disk or harrow, another to pull the fertilizer wagon and yet another tractor to pull the harvester. It was a practice that was wasteful in the extreme and all because there had been no vision, will or discipline to develop standards.

In 1994, as the work on the immigration project advanced, the director responsible wrote a discussion paper entitled *New Directions in Meeting Needs for Social Data*. [16] A section of that paper entitled *Vision for the Future* is included in Appendix A.

The paper argued that there was a need and an opportunity to press forward with data integration. It stated that first it was necessary to build a base of comprehensive meta-information that had to describe the content of microdata files, the content of aggregated tabular output, the content of analytical or descriptive reports and the nature of specialized services provided by the agency. The information need to be accessible through both keyword and thematic searches, ideally supported by a thesaurus.

Secondly, there had to be one gateway and one tool to access the meta-information, operating on-line and in real time. That is, clients should not have to contact the many divisions of the agency to find what they needed. They should be able to search on-line in real-time one source that would direct them to the information they needed.

Thirdly, the disharmony in the meta-information had to be addressed and resolved. One of the attributes of building a meta-information or metadata base is that it quickly reveals the considerable disharmony that exists across the various sources. The metadata base would provide a new tool for identifying this but the agency needed the will to address and resolve it.

Fourthly, there needed to be increased thematic outputs. Data and information should be released, not simply based on a single source as had been practiced in the past. It had to be integrated with other relevant data from all the agency's sources. The release of anything less than our comprehensive knowledge of an issue or population could seriously mislead the client.

[1] Steve Kerstetter, Acting Director, National Council of Welfare – letter to D. Bruce Petrie, Assistant Chief Statistician, Social Statistics, Statistics Canada. 24 August 1994.

Finally, the corporate culture had to change. This required initiative at the highest level of the agency. The failure to promote a shared vision, develop strategic planning and direction and provide funding sent the signal that integration might not really be a high and urgent priority.

It is difficult to say what impact the above-quoted paper had upon the agency. However, two events took place in 1995, the following year. The first was that the twelfth in a series of international symposia on methodological issues sponsored by Statistics Canada featured a session on data integration. Assistant Chief Statistician Gordon J. Brackstone, Informatics and Methodology Field, said in the opening remarks that "*...a survey should be thought of as contributing to a corporate base of information, which may contain data from many different sources, and from which information can be retrieved in a common integrated way – a corporate data base that provides the foundation for an information service utilizing all the data sets available, both singly and in combination. What this evolution reflects is the understanding that the results of a survey are not just a stand-alone set of tables, but an addition to an information base that may be used in many foreseen and unforeseen ways.*" [1]

Included in the program was a Statistics Canada-contributed paper called *Data Integration: the View From the Back of the Bus*. The abstract for the paper is as follows:

Statistical agencies have tended to be methods driven. That is, the collection activities took place through vehicles developed around specific methodologies. Each vehicle often served its own specialized clientele without regard to the needs of other organizations. The agency, therefore, often evolved, not as a corporation but a consortium, or even fragmented consortium, of relatively independent producers of data. Methods, systems, concepts, definitions, classifications, products and services were developed independently resulting in inefficiencies, redundancies, disharmonies and some client frustration.

The client satisfied with single-source information has been relatively well-served. But the client who needed comprehensive information on a particular issue, population or geography has not. Now information technology has precipitated a paradigm shift.

A new generation of clients is cutting its teeth on the Net and developing new expectations, particularly with respect to searches for information. These clients, all with their unique and particular needs, expect to be able to thematically browse meta information, determine sources, make selections and even download: on-line, real-time, seamlessly and at low or no cost.

The challenge to, and opportunity for, the statistical agencies is to respond to the new paradigm by accommodating these clients. The keystone to building such a response capability rests in integration. This includes both developing links between the sources and eliminating or reducing the disharmonies. Integration is also fundamental in moving from data to information because it facilitates bringing together all relevant and available inputs. Informed decision making depends on it. [18]

Perhaps the significance of the symposium was that the agency was formally acknowledging that culture had to change, that it had to move away from stove-pipe outputs based on isolated islands of production.

The second event of the year was the formation of a new directorate in the social statistics' field. It was charged with the integration and development of social statistics. With a small staff, a limited budget and a three-year time-frame, it was challenged to take what had been learnt from the building of the metadata base on immigrants and extend it to the whole social statistics' field. It was also to further the cause of integration and harmonization in the field.

New Directorate Committed to Integration, Harmonization And Metadata

The mandate for the new directorate was formulated and subsequently reported at the American Statistical Association meetings in Chicago in 1996. [20] It is quoted in part in Appendix 2.

While the paper repeated many of the messages that been laboured at many previous workshops, management conferences or committee meetings, this time it was being delivered on a respected international stage. On the one hand, it was an admission that Statistics Canada had become a consortium of unregulated enterprises using islands of technology or expertise that met specific needs but resulted in a fragmented organization. These islands had limited and specialized functions that sometimes had nothing to do with the overall mandate of the corporation. Furthermore, they had become balkanized with formidable physical and organizational barriers between them, leading to redundancies and inefficiencies. One of the legacies of this type of organization was lack of meta-information.

On the other hand, the paper clearly signalled a commitment to address these issues and move forward on both the fronts of integration and harmonization.

The paper went on to argue that:

Statistical agencies generally have 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 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.*

In addition to the lack of meta-information there was considerable disharmony between the various data sources:

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.

The paper noted that frequently a dataset from one source could not be compared with another source or that naming conventions for the same variable changed from source to source – or that classification systems varied amongst the sources. Furthermore, where there were independent vehicle-driven output data from different sources they might appear contradictory due to different strategies in rounding or seasonal adjusting. And, again, there was the matter of bias in single-source outputs. The release of a set of information from a single source without the benefit of related and relevant data from other existing sources could be dangerous. Partial data and therefore incomplete data could be misleading and lead to biased conclusions.

The paper reiterated points made in the earlier paper *Data Integration: the View From the Back of the Bus* and further stressed the need to build the metadata bases, address the disharmony, provide a single gateway and search tool on-line, real time, and develop integrated thematic outputs. It concluded:

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.

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 and 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.

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 opportunity provided by information technology will be quickly perceived as unhelpful and irrelevant.*

With this mandate, the new directorate began to build on its previously-developed but limited metadata base. The case might have been made to have the project work out of the agency's Standards Division, but for reasons that were never expressed or documented it was decided it would be conducted within the social statistics' field.

Building the Metadata System

The first issue was to address the problem of independent producers and the failed subject-matter committees. At that time there were approximately nine subject-matter divisions in the social statistics' field that made largely independent decisions about the data they collected. These directorates reported through four branches, each headed by a director-general who reported to the Assistant Chief Statistician (ACS) for Social, Institutions and Labour Statistics. The Director of the newly-formed Integration and Development of Social Statistics also reported directly to the ACS and as such attended the weekly executive meetings. It was at that level that direction was given to the program and compliance with the project enforced. One of the first activities was to start collecting metadata from the subject-matter divisions: definitions, classification systems, derivation algorithms, time-frames, questionnaires, data-base layouts, processing specifications, quality measures and other supporting documentation. An electronic template was developed that greatly facilitated the capture of any metadata already in electronic form. Initially there was resistance by a number of divisions. Some just were not prepared for the new culture, while others complained they had no resources to undertake the work and wanted the new directorate to pay for it. This was still a legacy of the user-pay philosophy introduced some years earlier. By executive order, the divisions were simply told they would have to absorb the cost in their existing budgets and get used to it, as this was the new world in which they would have to operate. It was surprising, once the priority given to the initiative by senior management was realized, how quickly the information began to flow. Once received, regardless of the format in which it had been stored, it was placed in a standard format in a hypertext database that permitted linkages not only between sources but also between components, such as definitions, classifications or products or outputs. The metadata were also organized into thematic entities such as health, labour, education, etc.

While the demand to develop metadata had increased, and the will to do so was now in place, perhaps the most important factor in the work actually proceeding was the storing of files and documents in electronic formats. New software was used that facilitated hypertext linking between the various documents. That made the endeavour both feasible and affordable. An attempt at Statistics Canada in the early 1980s (as reported to the Conference of European Statisticians [3]) to develop metadata on families failed because of the cost. At that time, with the technology available, it was estimated that it would take about three people a year to build the metadata and an only slightly lesser number to maintain them. Much as the desire was there to build the system, it just was not affordable. The new technology solved that problem.

Not to be overlooked are technological advancements that improved communications between the directors of the island or stove-pipe production areas. Meetings are not necessarily the most effective means of either communicating or decision-making. Telephone discussions often involve delays due to telephone tag and there is normally no record of what was discussed. Memos are cumbersome…Director One dictates it, his or her secretary types it, the director proofreads it, the secretary passes it to a mail clerk who delivers it. A second secretary assigns it a priority and, at some point, Director Two reads it. While today many might argue that electronic communication has overwhelmed us, the development of e-mail vastly improved communication in the agency. Communication became much faster and less formal and an electronic record was kept. All contributed to breaking down previous barriers.

When the social statistics' metadata project started, the agency had just succeeded in networking personal computers within the organization. Communication and the sharing of information blossomed accordingly. The agency Intranet was the perfect platform for lodging the social statistics' metadata base, which was named the Thematic Search Tool (TST). The database initially contained historical meta-information going back to 1984. It included over 125 statistical activities. Furthermore, the database could be searched on a number of parameters: year of collection, vehicle of collection, universes, variables, keywords or thematic subjects. Even while the meta-information was still being collected that which was already inputted and formatted was available for all staff to peruse. For the first time an officer in one shop could see the metadata not only from his or her division but from all the others as well. Suppose he or she was charged with developing a question and classification system for a variable that, for example, we might call 'ethnicity'. He or she could see immediately how that variable had been defined and classified in all other census, survey or administrative record sources. There was an immediate potential for significant cost savings in reducing developmental time. There was also an immediate potential for officers working in dissemination functions (such as the regional offices) to find data-sets that would be of use to their clients. And, at last, the stage was set to start the process of determining best practice and moving towards harmonized and standard definitions and classifications.

The latter, in fact, began to happen as new vehicles came on-line or as existing vehicles went into their next cycle. Officers responsible could not proceed until the proposed content, questionnaires, definitions, classification systems, etc, had been approved by the executive committee that reviewed all such requests with a view to determining standards and harmonized outputs. At long last, stove-pipes began to fall, isolated islands of production were bridged and the agency began to move from a consortium of independent producers to a corporate entity.

In less than two years from the launch of the initiative, the metadata base for social statistics, through the TST, was launched on the Internet through the agency's website. It had two immediate positive impacts. The first was that clients could almost instantly determine the extent of data holdings that might be applicable to their needs. The second was that those contemplating having the agency undertake special survey work for them could save developmental time by selecting variables, definitions and classifications already proven in earlier lines of inquiry.

In 1998, the project had completed its three-year mandate, funding dried up and the director in charge retired. The work continued to be updated but only until the last of the unfunded staff had been absorbed into other projects. While the benefits of the initiative continued to be felt, further progress was not realized until a few years later when senior management again made it a priority.

Shifts in Methods of Data Dissemination

Of course, the methods by which an agency disseminates its data are critical to the relationship it maintains with its clients. Throughout most of Statistics Canada's history, print publications were the primary if not sole method of dissemination. In fact, for many years the agency maintained its own printing plant. In order to assist data collection in such a vast country a number of regional offices were opened in 1945. By 1949 it was realized that they could also provide a useful role in dissemination and each office had a library where clients could access printed publications. Over the years the role of those libraries has expanded to provide consultation services as well. In 1965, CANSIM (the Canadian Socio-Economic Information Management System) was introduced as a data storage, retrieval and manipulation system. By 1972 its services were made available to clients on-line and today it provides ready access to data, updated daily, on a broad range of subjects and population groups.

Since the 1971 Census, public-use microdata files have been made available for every census. Also since the 1971 Census, a custom-tabulation service has been provided including outputs for special population groups and geographical areas as well as special customized mapping.

As technology has changed, print publications have diminished in importance as more data have been made available on-line or in DVD-ROM format. Browsing the Statistics Canada website provides direction to these services, most of which can be accessed for a fee. But there are also summary tables that can be downloaded at no cost.

As noted earlier, government constraints imposed in the 1980s forced the agency to move from the position that its products were a public good provided at no or low cost to a position that, if a product did not have a market value, it would not be produced. For example, the government decreed that the 1986 Census program would be responsible for "recovering costs of products and services (and) would generate…$44 million". [25] This notion of cost-recovery spread to all fields of the agency. It did have the benefit of reducing or eliminating marginal products. It also had the benefit of forcing a closer relationship with clients, producing products that were more useful and developing an aggressive marketing strategy. Major clients in government, industry and business, while complaining, generally adjusted to the new scenario and absorbed or passed on the new costs. Private citizens and small businesses, of course, simply could not absorb the new costs and were significantly disadvantaged, as were academics, teachers, students and researchers. Universities and other post-secondary institutions were very hard hit and Canadian research and teaching actually had to rely on foreign data.

Many of the advisory committees began to debate the problem in the course of their meetings and by 1989 the Canadian Association of Public Data Users and the Canadian Association of Research Librarians formed an *ad hoc*-buying consortium to gain access to microdata. By 1993 the Social Sciences Federation of Canada, assisted by Statistics Canada and the Depository Services Program, developed a proposal that was accepted by the Treasury Board of Canada. The outcome became known as the Data Liberation Initiative (DLI). It now permits 75 post-secondary institutions to offer a full range of free data services to students and faculty and academic researchers. Data libraries have become an integral part of these institutions. As such they have fostered Canadian research, improved teaching and developed a new generation of students who know how to use data in their studies and research. These new graduates are well trained, well informed and not satisfied with data or products that are not integrated, harmonized or useful. They represent a powerful force to ensure the agency does not reverse its push for standards.

The Last Stove-Pipe Falls; Metadata Come of Age

By 2000, however, it was realized that one more stove-pipe had to fall. The integration and harmonization initiative started in the social statistics' field needed to be absorbed into the Standards Division, where metadata from the whole agency could be managed under one roof and be accessible from one entry point. That system is well described in the report of Statistics Canada to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS).

In that paper the agency's Integrated Metadata base (IMDB) is described as follows:

Statistics Canada has implemented a corporate metadata base that stores metadata on its 566 current surveys and statistical programs. The IMDB contains another 312 records for survey in various states (e. g., surveys with no publicly disseminated data, amalgamated data, etc.) for historical purposes. The content of the IMDB has been selected to suit its primary purpose, which is to provide users with information needed to interpret the statistical data that Statistics Canada disseminates. The type of information provided covers the data sources and methods used to produce the data published from surveys and statistical programs, indicators of the quality of the data as well as the names and definitions of the variables, and their related classifications. The metadata supports all of the Agency's dissemination activities including its online data tables, CANSIM and Canadian Statistics, publications and daily data releases. The IMDB has been built to facilitate the maintenance of historical statistical metadata...[2]

Perhaps the best way to describe the IMDB is to direct readers to its public interfaces by visiting the Statistics Canada web-page.[3] The home page offers the option to *Find statistics* and *By subject*. Taking the subject option, statistics are listed by subject. Clicking on *Aboriginal peoples*, for example, reveals a number of sub-topics. Here, choosing *Aboriginal peoples (general)* reveals 20 publications, 30 analytical studies and nine links to definitions, data sources and methods. Here we find nine surveys or other sources of data on aboriginal peoples. Further drilling reveals the content of questionnaires and reporting guides, a description of the survey, data sources and methodology, data accuracy and the data files. We can also find definitions and classifications.

It should clearly be evident that metadata of this nature provide an indispensable tool in mining the vast archives of the statistical agency – archives previously held in dank, dark basements. Metadata are, indeed, the key to data liberation.

To quote again the invited paper:[4]

The IMDB is becoming the single source of metadata for describing Statistics Canada's surveys and statistical programs. This means that survey managers have to supply metadata on their individual surveys once to the stewards of the metadata, Standards Division. Since IMDB is built on a common metadata set with reusable administered items and attributes, survey managers can reuse descriptions for different survey cycles and across other surveys they might manage. Also, the use

[2] Invited paper submitted by Statistics Canada (2006), Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS), Geneva.

[3] www.statcan.gc.ca

[4] Invited Paper, op cit..

of a common metadata set presents a 'common look and feel' to data users accessing the metadata through our website.

While most of the content in the IMDB was determined by the Policy on Informing Users of Data Quality and Methodology[5] *both internal and external users have indicated other requirements when it comes to statistical metadata. The IMDB has been designed and continues to be developed to meet these needs. Now that the metadata is complete for every survey, other users can access those administered items that meet their requirements. In addition to supporting the information requirements for disseminated data, the IMDB is being used as a source of information for standardizing survey processes and content, corporate and financial planning, quality management at the survey level, survey respondents, international data exchanges and data researchers.*

Policy on Standards

Underlying the IMDB is Statistics Canada's Policy on Standards,[6] which states:

Statistics Canada aims to ensure that the information it produces provides a consistent and coherent picture of the Canadian economy, society and environment, and that its various datasets can be analyzed together and in combination with information from other sources.

To this end, the Agency pursues three strategic goals:

1. *The use of conceptual frameworks, such as the System of National Accounts, that provide a basis for consolidating statistical information about certain sectors or dimensions of the Canadian scene;*

2. *The use of standard names and definitions for populations, statistical units, concepts, variables and classifications in statistical programs;*

3. *The use of consistent collection and processing methods for the production of statistical data across surveys.*

The Policy in full is quoted in Appendix 3.

Both the standards policy and the creation of the IMDB represented significant progress in addressing the frustrations of data-users, both external and internal, but other work remained.

New Household Survey Strategy

In 2005 a number of working groups were struck with a view to addressing rising costs, increasing respondent resistance and changing client demands in the conduct of household surveys. In the preamble to the work it was noted that:

The new household survey strategy is in fact building on many initiatives already underway. What is really important in our present climate is the evidence of far more horizontal thinking, and willingness to see the challenges we face as shared. This is crucial to effecting change. The vision and priority setting must be established globally because no one division, branch or field can single-handedly make happen the changes envisioned here. [26]

The statement is profound for two reasons: one negative, one positive. Firstly, it is evident that even in 2005, after years of discussing the need for integration, the agency remained an unregulated enterprise with key divisions still working independently. Secondly, it is encouraging that once again a small group of brave souls was advocating change.

The working group called for all household surveys to be placed in an integrated collection platform, using standardized questionnaire modules and standardized processing specifications. It realized that a key component was the standardized questionnaire modules and that they required a top-down approach. That is, they recognized that the unregulated producers would not change unless senior management drove the process. They called for the creation of a senior management position to drive the work.

Subsequently, work did begin on an integrated collection platform that was to be used by all household surveys. Work also began on developing harmonized content that to date (2010) has seen the completion of 17 standardized question modules accompanied by programming specifications, standards' metadata and relevant notes. These are currently being used by internal survey developers with plans to make the information available to clients externally. The process of developing harmonized content started with working groups comprising subject-matter experts and survey

[5] Available on the Statistics Canada website (www.statcan.gc.ca)

[6] Policy on Standards (Revised 14 July, 2004, modified 2010-03-05), Statistics Canada (available on Statistics Canada website).

stakeholders who recommended, based in part on international and UN standards, harmonized concepts, definitions, classifications and metadata. Their work in turn was reviewed by a subject-matter review committee primarily comprising stakeholder directors. Advice was also sought from the external advisory committees noted elsewhere in this paper. When work was approved at this level it was recommended to the Directors' Committee for Harmonized Content. This group, chaired by a director-general, reviewed and commented upon the work but did not have the authority to approve or disapprove. Final approval came from the Methods and Standards Committee chaired by an assistant chief statistician. The 17 modules currently approved include close to 60 variables, primarily of a demographic and social nature.

It is now the policy of the agency to use harmonized content for both new questionnaires and questionnaires undergoing major modification. It is inevitable that some clients will request the use of non-standard questions or modules. Such requests are reviewed by the Subject Matter Review Committee, where they might be recommended or not recommended for exemption status (the decision being made by the senior Methods and Standards Committee). It has recently become clear that the drive for integration and harmonization is finally being led by senior officials and has become top priority for the organization.

Thus through the standards' policy, IMDB, the work on the integrated survey platform and the development of harmonized content, the agency is seen to be making progress on addressing the expressed frustrations of data users, both external and internal. Over the years there was much internal resistance to change but the determination of a relatively few individuals kept the initiative alive. The power of external influences, however, must be given due credit.

External Influences

As we have seen, the need for integration and harmonization was long recognized within the agency but we must also recognize external influences that eventually led to the current state. With regard to federal government departments or agencies, Canada Mortgage and Housing Corporation was a long-time advocate. In fact, they had been very successful in producing useful reports in which they integrated (as best they could) data from Statistics Canada from their own administrative records and administrative data from provincial governments and other sources. They were relentless, and quite properly so, in demanding that Statistics Canada improve its outputs. Other government departments became advocates only to varying degrees. Of course, when political pressure was applied, as was the case of Citizenship and Immigration and Indian and Northern Affairs, their interest was heightened.

The former Learned Societies of Canada, now known as the Canadian Federation for the Humanities and Social Science, is a large group of scholarly organizations that holds joint annual conferences. Members present and discuss academic papers on the latest work in their field. For many, their work is based on Statistics Canada output and as such they were well aware both of inadequacies in and problems of access to the data. They were generally very outspoken in their criticism of the agency and its seeming inability to improve the situation.

As we have discussed above, the media were particularly influential in drawing attention to both difficulties in dealing with the agency and attempting to use its outputs. Initially the relationship was combative but once agency culture changed and integration was embraced, and the focus shifted from data to information, a partnership was formed that led to their contribution of very positive and constructive advice to the agency.

Non-governmental organizations such as the CD Howe Institute, the National Council of Welfare, the Vanier Institute for the Family and the Canadian Institute for Health Information, as well as many others, were more than happy to provide pressure, encouragement and advice. In fact, the Canadian Institute for Health Information has very much become a model for data integration, as they provide integrated data from many sources of surveys and administrative records, analytical reports and workshops on data collection and quality.

The National Statistics Council, along with the Advisory Committees, especially the Advisory Committee on Social Statistics – which was particularly helpful in steering the development of the metadata base and the thematic search tool – were a powerful voice that advocated and defended expenditure on integration, metadata and harmonization.

Canada has a long history of working with the Organization for Economic Co-operation and Development, Eurostat and the Conference of European

Statisticians. In fact, Canadians served on the staff of the latter for many years. Officers from the social statistics' field participated for a time in the various work sessions, informal meetings and seminars of these organizations. Today one might easily reflect on the importance of all those work sessions in developing the internationally-harmonized data that underlay the founding of first the European Economic Community and then the European Union. Few if any Statistics Canada delegates ever returned from such meetings without a commitment to fighting for integration, metadata, standards and harmonization. Those meetings clearly demonstrated the need for the work, provided a platform for discussion of how it might proceed, and documented and encouraged compliance with best practice.

Lessons Learned

Over nearly 400 years, data gathering in Canada had generally been a single-purpose enterprise. Jesuit missionaries collected data to support their missions. Administrators, such as Jean Talon during the French regime, collected data to assist the demographic, social and economic planning of the colony. Subsequently, British administrators did the same. After confederation, federal government departments generally collected their own data, whether related to fisheries and oceans, agriculture or immigration. Even with the transfer of data-collection to a centralized agency. the single-purpose practice continued, sometimes based on statistical methodology, sometimes on subject-matter interest.

Isolated islands of interest and stove-pipe areas of production persisted, even though we have seen continual pleas for integration and harmonization. Certainly, a major turning point came in the late 1970s when the agency was perceived as irrelevant and archaic in the eyes of both the media and many major clients. It was apparent that independent production was wasteful of resources. Data were being under-utilized with independent rather than integrated production severely limiting potential exploitation of the data. It was clearly pointed out that independently releasing data from different sources at different times could very much contribute to misleading conclusions. Consider our journalist who viewed the three oranges as if each were an independent still-life. Integrated data permit a video-like view of the orange as it appears in its natural state, then as it is peeled and the lines of longitude are revealed, and then as it is sliced at the equator, revealing its wagon-wheel-like appearance. Nothing succeeds like success and once the agency began to increase its integrated outputs and shifted to producing information rather than just data, clients quickly perceived the utility and demanded even more.

As the integration of data proceeded, the legacies of stove-pipe production were increasingly revealed. There had to be harmonization of the concepts, definitions and classification systems. This would not be easy since many statistical series had maintained historical comparability for generations. Clearly, this was no time for the agency to be making unilateral decisions. Partnering with major clients through improving the consultation process, through the formation of the National Statistics Council and Advisory Committees, entering into discussions with the Learned Societies, as well as participating in international forums on standards – all helped move the agency towards finding compromise and embracing best practice.

It followed, once the commitment to harmonization was made, that the building of metadata bases was a requirement. Once these were populated by data from all the agency's sources, it facilitated three things. It provided the basis for developing standards, it significantly reduced resources needed in the development of new surveys, or updating existing surveys. Cutting and pasting from the metadata bases substantially curtailed the old practice of 'reinventing the wheel'. It also facilitated the much more effective mining of existing databases by clients, as they could now easily peruse the agency's data holdings and select the information they required. It provided a brave new tool for the true liberation of the data.

Given earlier attempts to integrate, harmonize and build metadata bases, it was abundantly clear that the process had to be driven from the top and the corporate culture had to change. A policy on standards had to promulgated, one single body had to be authorized to manage the process and one, and only one, single gateway to the metadata was to be permitted.

While the need for integration and harmonization had long been recognized, it is easy to ask the question 'Why did it take so long?'. In fairness, the technology was such that it would have been a difficult and costly undertaking. However, with the development of networked computers and powerful new software that permitted the storage and hypertext linking of records and data, the investment needed to develop metadata was reduced significantly. It was now imperative that the technology be embraced – there were no more excuses for inaction.

Finally, and perhaps the most difficult lesson learned, was the very thing that the minister of statistics from another country had warned, "Keep at arm's length from the government you serve. You must not only be non-partisan, objective, independent and relevant, but you must be seen to be non-partisan, objective, independent and relevant." Probably no one in the agency could have foreseen the unconstitutional announcement that a census had been cancelled. Of course, it was duly reinstated but not before considerable damage was done. It was learnt that it is imperative that the agency must maintain contact with the public it serves and it must be, and seen to be, independent, relevant and useful.

Challenges to Agencies in Developing Countries

The above-noted lessons in various ways all demonstrate forward-thinking initiatives to improve the utility of basic data. However, the challenge for participating countries will vary somewhat depending on whether they have an established history in the collection of social data through administrative records, censuses or household surveys; or whether they are relative newcomers to these activities. If the former, they might well be faced with the challenges of integration and harmonization that f faced Statistics Canada. If the latter, the job might be somewhat easier, although the lack of experience in using household surveys will be a challenge in itself. Nevertheless, both types of agencies will be challenged to develop metadata bases to assist in both the collection and dissemination of data. They will be challenged to develop standards – standards that will adhere to every degree possible with international standards. They will be challenged to integrate data from their various sources to ensure balanced and comprehensive reporting. And they will be challenged to provide information rather than just basic data.

Some of the keys to success in meeting these challenges will be the following:

National statistical agencies will need to carefully guard their political independence. Without doubt this can be a challenge but if the statistical outputs are to have credibility, and therefore utility, they must be, and seen to be, free of political interference. One of the best ways to ensure the independence of the statistical agency is to develop strong partnerships with major clients. That includes international agencies, government departments, academics, non-governmental agencies, business and the media.

Relationships with the media are especially important, not only from the perspective of listening to media and public perceptions of needed avenues of enquiry, but particularly with a view to providing the media with not just data but information. That is, analytical outputs, written in a journalistic style that can be easily pasted into media products. Information that the public can relate to, such as profiles of geographical regions or profiles of particular population groups, are useful. For example, profiles of lone-parent families, the aged, children and youth, or agricultural workers – all are products that help develop an informed population and provide critical information to decision-makers.

It is crucial that data from various sources within the agency are harmonized. Concepts, definitions, naming conventions and classifications – all must agree from one source to another. That means that agencies must waste no time in developing a policy on standards – a policy that embraces international standards as much as possible. The goals of the policy should ensure the use of conceptual frameworks for consolidating statistical information; the use of standard names and definitions for population groups, statistical units, concepts, variables and classifications; and the use of consistent collection and processing methods of production. The policy must be **controlled** and **enforced** at the **highest level** of the agency.

A metadata base must be developed to support the policy.

The main function of national statistical agencies is to produce and disseminate statistical data on the economic and social conditions in their country. Statistical data take the form of numbers of various types, in data files, statistical tables or in texts such as news releases and articles. These numbers on their own cannot be understood. This explanatory information is called metadata, and its presence is essential for the correct understanding and interpretation of statistical data.

At the most fundamental level, this explanatory information must cover at least the description of the data. A standard that is useful, and that is being used by Statistics Canada, to structure and present this type of metadata is ISO/IEC 11179 'Information Technology – Specification and Standardization of Data Elements.' In statistical terminology, data elements are commonly referred to as variables. This standard therefore provides a guideline for structuring and presenting basic descriptive information about variables. The *very process of creating descriptive information according to this standard, however, also has the effect of bringing about more consistency and rigor in the conceptualization, naming and organization of variables for which data are produced.* [11]

It might be added that experience has shown that metadata are best documented at the outset of any new survey design or redesign rather than after the fact. When they are part of the process, they actually assist it.

Metadata support three essential activities of statistical activity. First, metadata support the design and development of new surveys or the redevelopment of existing surveys. They provide an immediate record of what was done in the past and which might be used again in the future, thereby rendering significant efficiencies and cost savings. Metadata provide a platform for developing and maintaining standards and best practice, which also leads to efficiencies, cost savings, and improvements in the utility of the product. Finally, metadata also promote and encourage the effective mining of the agency's data holdings. Clients can easily browse the agency's full repository of data, both current and historical, and select what is most appropriate to their needs. The potential of the data archives is then more fully realized.

Data from various sources within the agency should be integrated or at least cross-referenced to other sources in order to provide as comprehensive a view as possible of any geographical region, population or issue. Failure to integrate data may provide an incomplete and therefore misleading understanding of a situation or population. It is natural and convenient to release the latest data from a census, any given survey or administrative source at the time of its availability; but providing references and links to related data from other sources is a 'must'. For example, in most societies, lone-parent families are a group at risk. To best understand the condition of such families, data are needed on family composition; labour-force activity and income of family members; educational attainment of adults or school attendance of children; health status of family members; and living conditions of the family. Typically, such a range of data generally comes from various sources and the analyst must be aware of, and have access to, such sources. Better still, if the agency has the resources, special profiles of such population groups should be produced. Other groups that might be at risk in most countries might be children and young people, the aged, or the unemployed. In some countries, it might be aboriginal people, immigrants, women or agricultural workers. In other countries. it might be geographical regions or inner-city areas that would benefit from such integrated profiles. However, it must be remembered that integration of data depends heavily upon how successful the agency has been in its harmonization and standardization initiatives.

Finally, there is the matter of getting data and information into the hands of clients and here we must distinguish between those who need information and those who need data. With respect to the former, we can include the whole population of the country. The special profiles mentioned above serve them well – if not directly, then indirectly through the media, who tend to use such studies extensively in reporting on the social condition of the country. Of course these kinds of integrated studies are also of immense value to relevant government departments, non-governmental organizations (NGOs) and educators.

Those who need data to conduct their own research include government departments, NGOs, businesses and academics. For these clients, the metadata are essential in order that they may easily conduct the searches to find the data they need for their work. For some, tabular data suffice; but increasingly they require microdata. Harmonization and standardization across databases are essential and, again, metadata are critical to ensure the most effective use of the outputs.

There is also the matter of the cost to the client of accessing the data. In some countries, the outputs and data files might be considered a public good and products might be distributed at little or no cost. In other countries, the statistical agency might be required to pass on some larger part of the cost, as was the case in Canada, where some clients were able to absorb the costs while others not. The aptly-named Data Liberation Initiative was a creative solution to put the data in the hands of academic users who otherwise would have been unable to access and use the data.

A country's statistical system needs to be viewed as a national resource, if not a national treasure. Its products and services must be accessible. The data are essential to sound economic, social and demographic planning of the country; and the same data, when transformed into information, are critical to the population's understanding of its social and economic condition and its place in the world. Developing countries that do not have multiple, long-standing statistical series or sources have a unique opportunity to start on the right foot, use the tools now available and, using standards, ensure their outputs are harmonized and integrated. Even for those that do have existing non-harmonized series, the sooner they commit to standards the less pain and disruption they will face over the longer term.

As can be seen in the forgoing narrative, the failure to integrate, harmonize and standardize has dogged Statistics Canada throughout its long history. Granted, with past technology, it was not an easy or affordable task. With the introduction of new technology and software in the 1980s and 1990s, the old excuses were no longer valid. The new challenge became the need to change the corporate culture. It was a long and sometimes painful path – as the agency followed up the papers, workshops and seminars of the '80s and '90s – before we saw action such as the recently-revised policy on standards and initiatives, e.g. the New Household Survey Strategy. Young statistical agencies in developing countries have an opportunity to learn from the experience of Statistics Canada: its long struggle for integration and harmonization provides a road map. *Those who cannot remember the past are condemned to repeat it*. [23]

Most of us have heard the expression, "Old soldiers never die; they just fade away". We might also say, "Old statisticians never die; they just no longer count". By extension, we might say that old-thinking statistical agencies that fail to integrate, harmonize and mine their data just no longer count!

Appendix 1

Quotation from Gordon E. Priest, *New Directions in Meeting Needs for Social Data*, Unpublished Discussion Paper, Statistics Canada, September 1994.

The vision of the future will see a clientele which is no longer content with standard on-the-shelf-take-it-or-leave-it products. Just as the introduction of the printing press marked the beginning of the liberation of the people from the dictates of the educated elite so will the introduction of personal computers, linked in global networks, be a great liberator. The playing field will become much more level when anyone armed with a personal computer, a modem and a minimal level of computer literacy will be able to access the collective knowledge of humankind. Encyclopaedias, great literature, digitized works of art and music, data and knowledge will be readily available and much of it will be free. The frustration expressed by sophisticated users like Bradley, other federal clients and academics, will be a mere whisper in comparison to what we will hear from the masses when they wish to browse our metadata and discover that there are none or they are not harmonized. We must give priority to developing standardized and harmonized metadata bases and making them easily accessible via CD-ROM and the Internet. We now have the technology to do it cheaply. We lack only the vision and the will.

Furthermore, most of our products serve a relatively few generic markets but we know that there are many, many niche markets which we cannot serve economically because they are so small. We must take advantage of the new technology to move to a common production platform for the products we do produce. If we moved to such a platform we would gain the ability to produce 'spin-off' products for niche markets that would require nothing more than a little client consultation and some electronic 'cutting and pasting' to produce new products. There is even the potential to allow clients electronic access to the production platforms…and to allow clients to peruse the information and do their own cutting and pasting. They could be charged on the basis of the information downloaded and we would avoid the cost of doing the consultation, compiling and production. In fact, allowing clients access to both metadata and real information in hypertext has the potential to lead to a significant cost reduction.

With regard to the issue of building a framework for social reporting or developing social indicators, the new environment offers opportunities to make some real progress. The very absence of metadata in the past limited the number of players, the potential for experimentation and the potential for discussion about what might be done. Only a select few had access to the kind of information needed for experimentation and they were generally preoccupied with the production of the data rather than with harmonization. Access to metadata and improved access to data will allow a much broader participation in the development and testing of models of social indicators and help resolve the problems of time-horizon, operationalization, replication and boundary definition noted by Carley. [2]

In addition, we have the opportunity to extend our reach to our clients well beyond the formally established consultative initiatives. The establishment of users groups with the Internet framework provides a forum for wide-ranging informal communication on areas of interest, be it based on issues, special populations or specific collection vehicles. For example, we are already exploring the possibility of establishing an Internet group for users of the General Social Survey through which they can share information about both the manipulation of the data bases and resultant research.

Tapscott and Caston suggest that the new enterprise must be open and that it needs to recast its external relationships. It must be prepared to serve a more knowledgeable, more empowered and demanding clientele. Michael Adams, president of Environics[7] noted that the market has become highly fragmented and that generic approaches are beginning to fail. The question is, how do we, in the face of stable or declining resources, shift from generic products to niche products?

The first step is to accept the fact that we have a number of very distinct market segments (and to a degree distinct mandates). For example, we have a mandate to ensure that the Canadian public is informed but that hardly constitutes a market. Nevertheless, in other cases mandates and markets may be synonymous. There is a need to provide information for government

[7] Environics Research Group is a private company that collects and analyzes data on consumers, financial services, health, public affairs, social values and cultural markets.

policy making and review, for government program development and evaluation, for supporting business decisions, for teaching and for research. That we have not served these markets as well as we might have should be evident. The answer lies in taking a corporate approach in dealing with them.

With regard to federal government departments we have had varying degrees of success in meeting their needs through interdepartmental committees but even in those cases persons at the working level may be left to their own devices to follow up with a multitude of contacts in our program areas or islands of production. To supplement or even replace some of the interdepartmental committees we might want to consider developing a cadre of account executives whose job would be to inform the client of the wide range of data and information available (or potentially available), do the liaison with the program areas and ensure that the clients needs are met. That would include ensuring that the most appropriate data set was provided or that appropriate integration of the data sets was undertaken depending on need. Given the support we receive from government departments and given their reliance on our products it is in our mutual interests to provide much better service. By extension the account executive approach may be taken with other sectors as well, particularly teaching and research.

Account executives, however, would need appropriate tools to do the job. As a minimum they would need to be armed with sound knowledge of the appropriated subject matter, sound knowledge of the various data collection programs and a current inventory of not only the full range of products and services, but the extent of the data holdings themselves...that is, a comprehensive electronic metadata base.

The electronic metadata base is fundamental, not only to supporting account executives, including those already operating in regional offices, but also in allowing any potential client to browse our data holdings. Furthermore, only in the building of such data bases will we ourselves learn the full extent of the conceptual, definitional and classification disharmonies that exist between them. Only then can we begin to actively address and resolve the disharmonies in a systematic way. Furthermore, the development of metadata bases will facilitate and improve our own integration and exploitation of the data, not only in the production of our current flagships but in developing new products as well.

A modest attempt has been made to develop an electronic inventory of questionnaires, an inventory of survey sources and an electronic inventory of products and services. But none of these activities can satisfy the need for shared knowledge of the extent and richness of our databases which include many derived variables not evident in examining an inventory of questionnaires and not evident in our inventory of pre-planned and catalogued products. In addition, we might expect that a better and common knowledge of our own data bases would improve our content and questionnaire development.

Beyond recognizing our segmented markets, providing account executives to serve them, developing better inventories and improving conceptual, definitional and classification harmonization there is the above-noted need to develop the common production platform to facilitate our meeting niche market needs and allowing niche markets to service their own needs.

We need to reconsider the development of a framework for both social indicators and social reporting. Clearly, much needs to be done on a sectoral basis in terms of improving the harmonization and integration of information (e.g., health, education, labour force). But beyond that there are two factors that might be considered which were not addressed in earlier attempts at developing social indicators.

The first is that the rationale for the development of social indicators was always expressed as the need for data which would facilitate policy formulation. It tended to assume clear distinctions between the activities of government, non-governmental (volunteer sector) and businesses. It did not recognize the degree to which the provision of various services (or goods in some cases) could be substituted between the three sectors. It tended to ignore that similar, if not the same information is needed for both business decisions and for policy and program development and monitoring by government and non-governmental organizations.

Thus, there is a need to develop series of social statistics which are common to the needs of the three sectors and which contribute to an understanding of the relationships between them.

The second shortcoming in earlier attempts at developing social indicators is that insufficient weight was given to family status. In most models the variable was recognized but marginalized. In fact, it should be central. The degree to which an individual consumes goods or needs services depends very much upon their family status. Consider the following and consider the degree to which an individual's relationship with the economy is influenced by his or her family status: expenditure and consumption, labour force participation, holding of assets and income. Income is often shared at a family level and assets are often held at a family level. Expenditure and consumption have family attributes and even one's labour force participation can be influenced by family status.

Finally, one's relationship with environmental considerations is strongly influenced by family status. Shelter obviously has shared family attributes as does concern for public security and physical setting (quality of neighbourhood). Health is strongly related to family status as it has been well- demonstrated that the course of any illness can be strongly influenced by the presence or absence of family support. Similarly, family relationships can be impacted heavily by the illness of one of its members.

Thus, the development of a framework for social reporting needs to place a much greater emphasis upon family status. And, as noted above, it needs to better recognize the sectoral relationships in the provision of services and goods. While many other initiatives suggested in this paper related to sectoral improvements in our output, some energy needs to also be directed to the search for a framework.

In summary, we need to recognize our segmented markets, initiate better mechanisms to serve them, take advantage of currently available technology to build and maintain metadata bases, improve harmonization, revisit the notion of a framework for social reporting, seek opportunities to integrate data and information, develop a standard production platform and most important of all empower our clients.

We ignore the new world order and the new technology at our peril. To quote Tapscott and Caston one last time, 'Organizations that do not make this transition will fail. They will become irrelevant or cease to exist.'[27]

The paper included a discussion on social indicators and social reporting to help demonstrate that, without integration and harmonization of sources of data, further talk was futile, as relationships are generally more powerful predictors than individual attributes.

Finally, the paper concluded with a section on actions to consider:

First in order to refocus our attention on an integrated corporate approach to marketing, we could explore the possibility of establishing a pilot for an account executive program (particularly with respect to servicing federal government departments). We have knowledge of the success of the Australian Bureau of Statistics in this regard. Furthermore, we have the model in some regional offices[8] and we have an opportunity through the Corporate Assignment Program[9] in obtaining someone to undertake the task. The activity could be placed in Advisory Services Division and monitored by that division, Marketing Division, Dissemination Division and the subject matter divisions which would benefit from the integration.

Second, pursue the matter of filling in the missing pieces of the social metadata base. Data Access and Control is building an electronic base which provides very summary data describing the nature of the various data sources. Standards Division is building a data base which provides some detail with regard to methodologies (and questionnaires) employed by the sources. Library Services is building a metadata base which describes the catalogued or registered products and service. The missing link, however, is a metadata base which describes the details of the universes and variables (many of which are derived) which are resident upon the operational data bases. This is the only base which can describe what can be either accessed through custom tabulation or provide the source for micro-data outputs. It needs to contain a full range of variables and their class intervals to be effective. Such a data base, when rendered key-word searchable, is essential to the work of subject matter officers, Advisory Services and it is essential if we are to open our doors more widely to clients through avenues such as the Internet…

1991 might serve as a base year. All divisions with social data holdings from, and including, that year should be required to submit, in a standard format, details about their data holdings complete with a listing of all universes and variables and their detailed classifications. This information should be consistent with the summary information already held by the other corporate players noted above.

[8] At the time, Statistics Canada had a number of regional offices across the country that assisted in both data collection and dissemination.

[9] An inter-departmental program that provided temporary work assignments in other departments whereby employees could broaden their work experience.

We are in the process of completing a metadata base on immigration data and this work might serve as a pilot. This exercise has clearly demonstrated how inconsistent we are in the documentation and description of our data bases. Nevertheless, the work will be finished before the end of the fiscal year. It covers many of the current sources of social data and it could serve the purpose of setting standards, which if used by all program areas in the future, will feed the metadata base automatically with no further investment.

Third, the subject matter committees should be charged with reviewing the metadata bases with a view to identifying inadequate documentation within any particular source and with identifying disharmonies between sources with respect to concepts, definitions and classifications. The committees need to be given a target for the completion of the work (once they have access to the metadata bases) and be provided with resources to undertake this detailed and complex work. This work should be monitored at the field level.

Fourth, data-producing programs should be charged with preparing adequate, standardized documentation on their data holdings and resolving disharmonies (other than those which are unavoidable due to methodological differences) and revising their programs accordingly. Programs should report on their progress in their annual reports.

Finally, Dissemination Division and Marketing Division, in concert with the relevant program divisions, should take the initiative to develop a standard production platform for output products. Dissemination should take the lead on the technical side while Marketing Division should take the lead in seeking opportunities to serve niche market. Consideration should be given to providing access to archived hypertext via the internet to allow clients to electronically cut and paste and build their own products with some sort of fee or royalty paid for the information downloaded.

Appendix 2

Quotation from Gordon E. Priest, *In Search of Data Integration: No Matches Found,* Proceedings of the American Statistical Association, Volume 1, Chicago, 1996.

Statistical agencies generally have 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 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 are actually 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 variable 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. These are 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 Outputs 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 is 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 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 holding, 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.

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 of 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 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 you, 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 that 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-time, 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 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 are now 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 layout, 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 system 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 desktop. 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 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 measure 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 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.

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 and geography, and attempts to convert those data to information, has 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 opportunity provided by information technology will be quickly perceived as unhelpful and irrelevant.

Appendix 3

Statistics Canada Policy on Standards, revised July 14, 2004.[10]

Introduction

Statistics Canada aims to ensure that the information it produces provides a consistent and coherent picture of the Canadian economy, society and environment, and that its various datasets can be analyzed together and in combination with information from other sources.

To this end, the Agency pursues three strategic goals:

1. The use of conceptual frameworks, such as the System of National Accounts, that provide a basis for consolidating statistical information about certain sectors or dimensions of the Canadian scene;

2. The use of standard names and definitions for populations, statistical units, concepts, variables and classifications in statistical programs;

3. The use of consistent collection and processing methods for the production of statistical data across surveys.

This policy deals with the second of these strategic goals. It provides a framework for reviewing, documenting, authorizing, and monitoring the use of standard names and definitions for populations, statistical units, concepts, variables and classifications used in Statistics Canada's programs. Standards for specific subject-matter areas will be issued from time to time under this Policy as required.

Policy

Statistics Canada aims to use consistent names and definitions for populations, statistical units, concepts, variables and classifications used in its statistical programs. To this end:

1. Statistical products will be accompanied by, or make explicit reference to, readily accessible documentation on the definitions of populations, statistical units, concepts, variables and classifications used.

2. Wherever inconsistencies or ambiguities in name or definition are recognized between related statistical units, concepts, variables or classifications, within or across programs, the Agency will work towards the development of a standard for the statistical units, concepts, variables and classifications that harmonize the differences.

3. Standards and guidelines covering particular subject-matter areas will be issued from time to time and their use will be governed by the provisions of this Policy.

4. Where departmental standards have been issued, program areas must follow them unless a specific exemption has been obtained under the provisions of this Policy.

5. Programs should, to the extent possible, collect and retain information at the fundamental or most detailed level of each standard classification in order to provide maximum flexibility in aggregation and facilitate retrospective reclassification as needs change.

6. When a program uses a population, statistical unit, concept, variable or classification not covered by a departmental standard, or uses a variation of a standard approved as an exemption, it shall use a unique name for the entity to distinguish it from any previously defined standard.

7. Clients of Statistics Canada's consultative services should be made aware of and encouraged to conform to the standards and guidelines issued under this Policy.

8. The Agency will build up a database of names and definitions used in its programs and make this database accessible to users and other players in the statistical system.

Scope

This policy applies to disseminated data however collected, derived or assembled, and irrespective of the medium of dissemination or the source of funding. The policy may also be applied to data at the stage of collection and processing at Statistics Canada.

[10] See www.statcan.gc.ca.

Guidelines for the Development and Documentation of Standards

A. Introduction

These guidelines describe the requirements and give guidance for the development and documentation of standard names and definitions of populations, statistical units, concepts, variables and classifications. Section B defines the terminology; guidelines follow in Section C.

B. Terminology

For purposes of these guidelines the following terms are used.

Population: The set of statistical units to which a dataset refers.

Concept: A general or abstract idea that expresses the social and/or economic phenomenon to be measured.

Statistical Unit: The unit of observation or measurement for which data are collected or derived. The following list provides examples of standard statistical units that have been defined.

Person
Census family
Economic family
Household
Dwelling
Location
Establishment
Company
Enterprise

Variable: A variable consist of two components, a statistical unit and a property. A property is a characteristic or attribute of the statistical unit.

Classification: A classification is a systematic grouping of the values that a variable can take comprising mutually exclusive classes, covering the full set of values, and often providing a hierarchical structure for aggregating data. More than one classification can be used to represent data for a given variable.

Example:

The following is an example of the variable: Age of Person.

Concept: Based on the subjects used by Statistics Canada to organize its statistical products and metadata, the variable Age of Person is listed under the concept of Population and Demography.

Statistical unit and property: the statistical unit and property that define this variable are social statistics programmes. Age refers to an individual – this is the unit of analysis for most social statistics programmes. Age refers to the age of a person (or subject) of interest at last birthday (or relative to a specified, well-defined reference date).

Classification: Different classifications can be used to represent data for this variable. These classifications include: Age Categories, Five-year Age Groups; and Age Categories, Life Cycle Groupings.

The standard names and definitions of populations, statistical units, concepts, variables and classifications will be stored in the Integrated Metadatabase (IMBD). In the case of variables, the name stored in the IMBD will include a representation type, in addition to the statistical unit and property. In the age example given here, the full name of the variable in the IMBD would be Category of Age of Person. The representation type Category indicates that it is a categorical variable, which will be represented by a classification of age groups.

C. Guidelines

Each standard should have the following characteristics:

* describe the concept that the standard addresses when appropriate;

* identify the statistical unit(s) to which it applies;

* provide a name and definition of each variable included in the standard;

* provide the classification(s) to be used in the compilation and dissemination of data on each variable.

The most detailed level of a classification will always be included in a standard. Recommended and optional aggregation structures may also be present.

Concepts shall be described in relation to a framework where possible.

Every variable shall be given a name, in both official languages, which, once given, cannot be used to denote any other variable. Variables shall be defined with explanatory notes in terms of a property and the statistical unit to which it applies. Additionally, in the IMBD, the representation type will be defined.

Every Classification shall be given a name, in both official languages, which, once given, cannot be used to denote any other classification. Classifications shall be defined, with exclusions listed and explanatory notes given, where required.

Every class shall be given a name, in both official languages, which, once given, cannot be used to denote any other classification. Classifications shall be defined, with exclusions listed and explanatory notes given, where required.

The most frequently used populations shall be given a name, in both official languages, which, once given, cannot be used to denote any other population. These populations shall be defined with explanatory notes.

Every statistical unit shall be given a name, in both official languages, which, once given, cannot be used to denote any other statistical unit. Statistical units shall be defined with explanatory notes.

A standard shall be accompanied by a statement of conformity to relevant internationally recognized standards, or a description of the deviations from such a standard and, where possible, a concordance with the referenced standard.

Where a standard replaces an earlier one, a concordance between the old and the new shall be give.

A standard shall include a statement regarding the degree to which its application is compulsory. The different degrees are, in descending order of compulsion:

* **departmental standard:** a standard that has been approved by the Policy Committee, and the application of which is therefore compulsory, unless and exemption has been explicitly obtained under the terms of this policy;

* **recommended standard:** a standard that has been recognized by the Methods and Standards Committee as a recommended standard, with or without a trial period of a specified duration, after which it may be declared as a departmental standard;

* **program-specific standard;** a standard adopted by a statistical program, and which is registered with Standards Division, to ensure consistency in a series over time periods.

References

[1] Brackstone, G. J. 1995. “Proceedings of Statistics Canada Symposium 95: From Data to Information – Methods and Systems”. Statistics Canada.

[2] Carley, M. 1981. “Social Measurement and Social Indicators”. London. George Allen & Unwin, pp.87-111.

[3] Conference of European Statisticians.1983. “Report on the Problems of Integrating Household and Family Data from Different Sources, Different Concepts or Different Definitions of the Same Concept”. Submitted by Statistics Canada at Informal Meeting on Statistics of Households and Families, Geneva, 12-14 January 1983.

[4] Conference of European Statisticians .1986. “Considerations for the Definitions and Classification of Households and Families and Related Variables for the 1990 Round of Censuses”. Submitted by Statistics Canada, Geneva, 22-24 September 1986.

[5] Conference of European Statisticians. 1996. “Challenges and Opportunities in Administrative Records”. Submitted by Statistics Canada, ECE/Eurostat Joint Work Session on Administrative Registers and Administrative Records in Social and Demographic Statistics, Luxembourg.

[6] Conference of European Statisticians. 1996. “Issues of Meta Information and Integration”. Submitted by Statistics Canada, ECE/Eurostat Joint Work Session on Administrative Registers and Administrative Records in Social and Demographic Statistics, Luxembourg.

[7] Conference of European Statisticians. 1998. “Report on Progress on the Harmonization of Social Statistics”. Submitted by Statistics Canada at Work Session on Statistical Metadata, Geneva.

[8] Conference of European Statisticians. 2006. “Metadata to Support the Survey Life Cycle”. Submitted by Statistics Canada, Joint UNECE/Eurostat/OECD Work Session on Statistical Metada (METIS), Geneva.

[9] Fellegi, I. P. 1995. “Characteristics of an Effective Statistical System”. Morris Hansen Lecture, Washington Statistical Society.

[10] Gillman, D. W. 2005. “Common Metadata Constructs for Statistical Data”. Proceedings of Statistics Canada Symposium: Methodological Challenges for Future Information Needs.

[11] Johanis, P. 2005. “Documenting Data Elements in Statistical Agencies, Proceedings of Statistics Canada Symposium, 2005: Methodological Challenges for Future Information Needs”. Statistics Canada.

[12] Mechanda, K., Paul Johanis and Michael Webber. 2005. “Conceptual Model for the Definitional Metadata of a Statistical Agency”. Proceedings of Statistics Canada Symposium: Methodological Challenges for Future Information Needs.

[13] Nordbotten, S. 1993. “Statistical Meta-Knowledge and Meta-Data”. Presented at the Workshop on Statistical Metadata Systems, Eurostat, Luxembourg.

[14] Priest, G. E. 1989. “Views on the Terminology, Concepts, Indicators and Statistical Series on the Social Situation of Families”. Submitted by Statistics Canada, United Nations Inter-Regional Seminar, Yalta, 5-15 December 1989.

[15] Priest, G. E. 1994. “Social Information: Closing the Circuit”. Mountain West Canadian Studies Conference, Simon Fraser University, Vancouver.

[16] Priest, G. E. 1994. “New Directions in Meeting Needs for Social Data”. Unpublished Discussion Paper. Statistics Canada.

[17] Priest, G. E. 1995. “Data Integration: The View From the Back of the Bus”. Proceedings of Statistics Canada Symposium 95: From Data to Integration.

[18] Priest, G. E. 1995. “Proceedings of Statistics Canada Symposium 95: From Data to Information – Methods and Systems”.

[19] Priest, G. E. 1995. “The Family as a Unit of Analysis in Social Indicators”. Unpublished discussion paper. Statistics Canada.

[20] Priest G. E. 1996. “In Search of Data Integration: No Matches Found”. American Statistical Association Proceedings of the Section on Survey Research Methods, Vol 1, pp. 40-44.

[21] Priest G. E. 1996. “Challenges and Opportunities in Administrative Records”. Presented paper, American Statistical Association, Chicago.

[22] Priest G. E. 1996. “The Issue of Harmonization of Data from Diverse Sources”. Presented paper, Eurostat Workshop, London.

[23] Santayana, George. 1905. “Life of Reason, Reason in Common Sense”. Scribner’s.

[24] Statistics Canada. Report on the Problems of Integrating Household and Family Data

from Different Sources, Different Concepts or Different Definitions of the Same Concept. Submitted at Informal Meeting on Statistics of Households and Families, Conference of European Statisticians, Geneva, 12-14 January 1983.

[25] Statistics Canada. 1993. “75 Years and Counting: A History of Statistics Canada”

[26] Statistics Canada. 2005. Unpublished working paper, New Household Survey Strategy.

[27] Tapscott, D., & Caston, A. 1994. “Paradigm Shift: The New Promise of Information Technology”. New York. McGraw-Hill.

[28] United Nations. 2009. “Statistical Metadata in a Corporate Context: A Guide for Managers”. Geneva.

[29] Vozel, K. 1993. “Technical Evolution White Paper. Discussion Paper”. New York. AT&T.

About the IHSN

In February 2004, representatives from developing countries and development agencies participated in the Second Roundtable on Development Results held in Marrakech, Morocco. They reflected on how donors can better coordinate support to strengthen the statistical systems and monitoring and evaluation capacity that countries need to manage their development process. One of the outcomes of the Roundtable was the adoption of a global plan for statistics, the Marrakech Action Plan for Statistics (MAPS).

Among the MAPS key recommendations was the creation of an International Household Survey Network. In doing so, the international community acknowledged the critical role played by sample surveys in supporting the planning, implementation and monitoring of development policies and programs. Furthermore, it provided national and international agencies with a platform to better coordinate and manage socioeconomic data collection and analysis, and to mobilize support for more efficient and effective approaches to conducting surveys in developing countries.

The IHSN Working Paper series is intended to encourage the exchange of ideas and discussion on topics related to the design and implementation of household surveys, and to the analysis, dissemination and use of survey data. People who whish to submit material for publication in the IHSN Working Paper series are encouraged to contact the IHSN secretariat via info@ihsn.org.

www.ihsn.org
E-mail: info@ihsn.org